Investigating the relation between gravitational wave tests of general relativity

Nathan K. Johnson-McDaniel,[1,2] Abhirup Ghosh,[3] Sudarshan Ghonge,[4] Muhammed Saleem,[5,6] N. V. Krishnendu,[7,8] and James A. Clark[9,4]

[1]*Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom*
[2]*Department of Physics and Astronomy, University of Mississippi, University, Mississippi 38677, USA*
[3]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), D-14476 Potsdam-Golm, Germany*
[4]*Center for Relativistic Astrophysics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*
[5]*School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA*
[6]*Chennai Mathematical Institute, Siruseri 603103, Tamil Nadu, India*
[7]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Callinstrasse 38, D-30167 Hannover, Germany*
[8]*Leibniz Universität Hannover, D-30167 Hannover, Germany*
[9]*LIGO, California Institute of Technology, Pasadena, California 91125, USA*

(Received 20 September 2021; accepted 22 December 2021; published 9 February 2022)

Gravitational wave observations of compact binary coalescences provide precision probes of strong-field gravity. There is thus now a standard set of null tests of general relativity (GR) applied to LIGO-Virgo detections and many more such tests proposed. However, the relation between all these tests is not yet well understood. We start to investigate this by applying a set of standard tests to simulated observations of binary black holes in GR and with phenomenological deviations from GR. We consider four types of tests: residuals, inspiral-merger-ringdown consistency, parametrized phasing (two versions), and modified dispersion relation. We also check the consistency of the unmodeled reconstruction of the waveforms with the waveform recovered using GR templates. These tests are applied to simulated observations similar to GW150914 with both large and small deviations from GR and similar to GW170608 just with small deviations from GR. We find that while very large deviations from GR are picked up with high significance by almost all tests, more moderate deviations are picked up by only a few tests, and some deviations are not recognized as GR violations by any test at the moderate signal-to-noise ratios we consider. Moreover, the tests that identify various deviations with high significance are not necessarily the expected ones. In particular, the parametrized tests recover PN test parameters much closer to zero than their true values in some cases. Additionally, we find that of the GR deviations we consider, the residuals test is only able to detect extreme deviations from GR that no longer look like binary black hole coalescences in GR. The reconstruction comparison shows more promise for detecting relatively small GR deviations in an unmodeled framework, at least for high-mass systems.

DOI: 10.1103/PhysRevD.105.044020

I. INTRODUCTION

Gravitational wave (GW) observations have provided our first probes of the dynamics of general relativity (GR) in the strong field, highly dynamical regime. A variety of tests were applied to the first detection [1] and other detections in the first observing run of the advanced GW detectors (O1) [2] and more tests have been added as subsequent detections have been analyzed by the LIGO and Virgo collaborations [3–11]. These observations have so far revealed no inconsistencies with the predictions of GR and future observations with upgraded and new detectors [12–18] will place even more stringent bounds on any possible deviations from GR's predictions, or potentially reveal more subtle deviations that have eluded detection to date. These improved results will come both from more sensitive observations of individual signals, as well as by combining together many detections. See, e.g., [19] for predictions of the sensitivity of tests of GR with future detectors.

However, the tests so far applied by the LIGO and Virgo collaborations are all null tests of one sort or another—none of them is testing a specific alternative theory. There are ways of mapping the results of some of these tests to constrain specific alternative theories, e.g., using the results of the parametrized tests that vary post-Newtonian (PN) coefficients [20–24]. However, these mappings require

significant assumptions that may not be valid in practice. This is due in part to the relatively low PN order and linearization in the coupling constant of the calculations used in the mapping. However, perhaps even more importantly, the LIGO-Virgo constraints on deviations in individual PN coefficients from their GR values that are mapped onto the constraints on modified theories are not necessarily valid constraints on those PN coefficients in a theory where multiple PN coefficients vary, and/or when the merger-ringdown phase is also modified, as we shall see here. The LIGO-Virgo analyses are designed to detect deviations from GR, not to measure individual PN coefficients. As discussed in [25], tests against specific alternative theories are preferable, and there is preliminary work on such constraints in [26,27], though still with simplified waveform models.

In fact, numerical simulations in alternative theories are not yet quite advanced enough to provide even single waveforms that could be used to check the performance of the tests carried out to date, though such simulations are progressing (see, e.g., [28–34] for binary black hole simulations and [35–37] for simulations of binary neutron stars), as are analytical calculations (e.g., [21,26,38–49]). Additionally, while there is progress in simulations that could be used as proxies for non-GR effects, e.g., of charged binary black holes [50,51], and of binaries of black hole mimickers, such as boson stars [52–59], these are also not quite advanced enough to provide waveforms of the quality required for data analysis.

Given this, as well as the continued proposals for new null tests of GR or of binary black hole nature (e.g., [60–77]), it is important to understand the relation between the tests being applied to the data, to decide which set of tests is most efficacious at detecting a range of plausible deviations from GR and how a deviation from GR would show up in the various tests. For the present study, we restrict ourselves to the four waveform-based tests applied to the binary black hole signals through O2 [7] (with updated results for more events in [10,11]), namely the residuals test; the inspiral-merger-ringdown (IMR) consistency test; the parametrized test of GW generation, both the test infrastructure for general relativity (TIGER) and flexible theory-agnostic (FTA) varieties; and the parametrized test of propagation. We will compare the unmodeled waveform reconstructions with the waveforms inferred from the modeled GR analysis, as carried out in [78–80].

To do this, we apply the tests to simulated observations of waveforms with parameters similar to GW150914 [81] and GW170608 [82], as a paradigmatic high- and low-mass event, respectively. We also apply the tests to simulated observations of non-GR waveforms of the type used in the parametrized tests (purely phenomenological in the case of the tests of GW generation, and coming from the dispersion due to a massive graviton for the GW propagation test), as well as the self-consistently modified effective-one-body (EOB) waveforms used to check the performance of the IMR consistency test in [83,84].

We give an overview of the tests considered in Sec. II, describe the specifics of the simulated observations in Sec. III, present the results of the tests in Sec. IV, and conclude in Sec. V. We give the two-dimensional (2D) IMR consistency test plots in the Appendix.

II. TESTS OF GR

Here we give an overview of the tests of GR we consider in this study. All these tests rely on accurate models for binary black hole waveforms in GR, for which we primarily use the IMRPhenomPv2 model [85–87] as in all but the latest LIGO-Virgo tests of GR with binary black holes (e.g., [1,7,10]). IMRPhenomPv2 models gravitational waves from black hole binaries on quasicircular orbits including the leading effects of spin precession. The LIGO-Virgo analyses also use the SEOBNRv4_ROM model [88] to give a check of the effects of waveform systematics, though it only allows for nonprecessing spins. Here we only use SEOBNRv4_ROM for the FTA parametrized test, since this is the only test for which that model is used to obtain the primary results in the LIGO-Virgo analyses.

The analysis of GW data is often carried out in the framework of Bayesian inference and this framework is also used for all these tests in some form. In particular, to sample the likelihood, we use the implementation of nested sampling [89] in the LALInference code [90], which is part of the LIGO Scientific Collaboration Algorithm Library Suite (LALSuite) [91]. We compute the likelihood integral from a low frequency of 20 Hz to the Nyquist frequency of 1024 Hz, as used in the LIGO-Virgo tests of GR for GW150914 and GW170608 (the two events we use as models for our simulated observations) in [7].[1] Additionally, two tests also use the BAYESWAVE code [92,93], which uses Morlet-Gabor wavelets to model the gravitational wave signal, as opposed to a waveform model based on GR.

The first two tests described below use BAYESWAVE to check the consistency of the waveforms inferred from the data using a GR model and LALInference. In one test we subtract the best-fit GR waveform inferred by LALInference from the data and use BAYESWAVE to compute the residual signal-to-noise ratio (SNR). For the second test, we use BAYESWAVE to reconstruct the waveform directly from the data and compare the overlap of the reconstructed waveform with the GR waveforms that the LALInference analysis finds are good fits to the data. The third test we consider also tests the consistency of the signal, this time of the low- and high-frequency portions. Both of these portions are

[1]The analysis of GW170608 required a larger minimum frequency of 30 Hz in the LIGO Hanford detector, due to the detector state at the time [82]. We have used a low frequency of 20 Hz in all detectors.

used to infer the final mass and spin and these two inferences are then checked for consistency. The next pair of tests checks that various parametrized deviations from the GR waveform model are consistent with their GR value of zero—the two tests differ in how these parametrized deviations are introduced. The final test we consider introduces a parametrized dispersion relation and constrains deviations from the nondispersive propagation of GWs in GR. Henceforth, we use the following abbreviations to refer to the various tests: IMR: inspiral-merger-ringdown consistency test; TIGER: test infrastructure for general relativity parametrized test; FTA: flexible theory-agnostic parametrized test; MDR: test of the modified dispersion relation. We now describe these tests in full detail.

A. Residuals

Ground based GW detectors are characterized by noise from various sources in different parts of the frequency spectrum (see, e.g., [94,95] for a discussion of noise sources in the Advanced LIGO detectors). The noise in the detector in any given time is generally assumed to be stationary Gaussian noise colored by the detector power spectral density (PSD). To analyze a potential GW candidate event, we model the detector time series as a summation of signal and noise (see, e.g., [96]),

$$\mathbf{d}(t) = \mathbf{h}(t) + \mathbf{n}(t). \tag{1}$$

Here $\mathbf{d}(t)$ is the detector data, $\mathbf{h}(t)$ is the GW model waveform, and $\mathbf{n}(t)$ is Gaussian noise. The boldface notation here is used to specify that the quantities are vectors with one component each for every detector in the ground-based GW detector network. We infer the best-fit (maximum likelihood) waveform, $\mathbf{h}_{\mathrm{maxL}}(t)$, using LALInference and a GR model waveform, here IMRPhenomPv2. If $\mathbf{h}_{\mathrm{maxL}}(t)$ is an accurate estimate of the true signal, then the residual, defined as $\mathbf{r}(t) = \mathbf{d}(t) - \mathbf{h}_{\mathrm{maxL}}(t)$, should be consistent with noise. We test this consistency by analyzing $\mathbf{r}(t)$ with BAYESWAVE. Since BAYESWAVE relies on wavelets to model the signal waveform, any loud multidetector coherent features in $\mathbf{d}(t)$ not accounted for by the signal model $\mathbf{h}(t)$ are potentially reconstructed as parts of the signal reconstruction. In the case of a faithful reconstruction of the true underlying signal by $\mathbf{h}_{\mathrm{maxL}}(t)$, the BAYESWAVE signal model will produce waveform samples whose median is consistent with noise.

Similar to [7,10,11], we constrain the loudness of the residual by calculating the 90% credible upper limit on the network SNR ρ_{res} of the waveform samples. For the case of Gaussian noise, this tends to be $\lesssim 5$ for the LIGO-Hanford, LIGO-Livingston, and Virgo network. Specifically, we generated 200 sets of simulated Gaussian noise time series in each detector colored with the same "O3low" detector PSDs [12] used in this analysis, and analyzed them with BAYESWAVE. We then computed the 90% credible upper limit on the network SNR on each of these observations and found that 90th percentile of this distribution is ~ 5.

B. Reconstructions

The residuals test is used to place constraints on the quality of the signal reconstruction by characterizing the residual and studying its consistency with background noise. The waveform reconstructions test on the other hand, approaches the question of signal consistency by studying the waveform itself. LALInference and BAYESWAVE both offer signal reconstructions, $\mathbf{h}_{\mathrm{LI}}$ and $\mathbf{h}_{\mathrm{BW}}$, respectively. Both these algorithms rely on fundamentally different waveform models, i.e., GR-based and wavelet-based respectively. An agreement between their signal reconstructions gives support to the GR model used in the LALInference reconstruction. We quantify this agreement by computing the overlap between the two waveforms. The overlap is defined as the noise weighted inner product of two normalized signals (discussed further in, e.g., [96]). In our case, following [97], we define

$$\mathcal{O}_{\mathrm{B,L}} \coloneqq \frac{\langle \mathbf{h}_{\mathrm{LI}} | \mathbf{h}_{\mathrm{BW}} \rangle}{\sqrt{\langle \mathbf{h}_{\mathrm{LI}} | \mathbf{h}_{\mathrm{LI}} \rangle \langle \mathbf{h}_{\mathrm{BW}} | \mathbf{h}_{\mathrm{BW}} \rangle}}, \tag{2}$$

where $\langle \cdot | \cdot \rangle$ applied to boldface quantities indicates an inner product taken over the network, defined by

$$\langle \mathbf{a} | \mathbf{b} \rangle \coloneqq \sum_{i=1}^{n} 4\mathrm{Re} \int_0^\infty \frac{\tilde{a}^i(f) \tilde{b}^{i*}(f)}{S_n^i(f)} df. \tag{3}$$

Here n is the number of detectors (3 in the cases we consider) and the superscript i is used to denote the signal in the ith detector, whose PSD is $S_n^i(f)$. $\tilde{a}^i(f)$ is the Fourier transform of the time series a^i, and the superscript $*$ denotes the complex conjugate. Dividing by the PSD makes the overlap most sensitive to differences in the waveforms at frequencies where the detectors are most sensitive. The absolute value of the overlap is bounded between 1 (complete agreement) and 0 (complete disagreement). The overlaps for GW150914 and GW170608 are ~ 0.98 and ~ 0.58, respectively (see Fig. 2 in [97]). The much smaller overlap for GW170608 is due to its smaller SNR and particularly its lower mass, which spreads out the power in the signal over a longer time and makes it more difficult for BAYESWAVE to reconstruct the signal accurately. We expect (and find) larger overlaps for the simulated GR cases with no noise considered here, since both reconstructions are made less precise by noise, particularly the non-Gaussian noise that is actually present in gravitational wave detectors.

Since the values of the overlaps we find are generally quite close to 1, we present results in terms of

$\bar{\mathcal{O}} \coloneqq 1 - \mathcal{O}_{\mathrm{B,L}}$. We compute $\bar{\mathcal{O}}$ for the entire distribution of $\mathbf{h}_{\mathrm{LI}}$ with the median $\mathbf{h}_{\mathrm{BW}}$ waveform and obtain a distribution on $\bar{\mathcal{O}}$. We use a point estimate for $\mathbf{h}_{\mathrm{BW}}$ (i.e., the median waveform) since individual BAYESWAVE waveforms samples do not represent a physical waveform, but their median does represent a physically stable estimate of the true waveform. More details about this choice can be found in [97].

C. IMR consistency test

A binary black hole coalescence goes through three distinct phases: an initial *inspiral* where the two black holes spiral in due to the backreaction from GW emission, a *merger* where the two black holes coalesce to form a single remnant object, and a final stage of *ringdown* where the remnant black hole settles into a stable Kerr configuration through the emission of a quasinormal-mode spectrum of gravitational waves. Within the stationary phase approximation, the low- (high-)frequency portion of the frequency domain gravitational-wave signal frequencies comes from the early (late) portion of the time domain signal (see, e.g., the illustration in Fig. 10 of [84]). Thus, one can test the consistency of the inspiral and merger-ringdown portions of the signal by checking the agreement of the low- and high-frequency portions of the signal. Specifically, we choose to split the analysis at a frequency f_{cut} given by the (redshifted) frequency of the innermost stable circular orbit corresponding to the remnant black hole [98]. One can then use these mutually exclusive parts of the signal to obtain two independent measurements of the initial masses and spins and then apply analytical fits to numerical relativity simulations [99–101] to these quantities to infer independent estimates of the mass and spin of the final black hole, $(M_{\mathrm{f}}, \chi_{\mathrm{f}})$.[2]

The inspiral-merger-ringdown (IMR) consistency test checks that these two independent estimates of the final mass and spin are consistent with each other, as they must be if the data is well described by the waveform model used to perform this inference. We thus define fractional deviations in the estimates of the final mass and spin,

$$\frac{\Delta M_{\mathrm{f}}}{\bar{M}_{\mathrm{f}}} \coloneqq 2\frac{M_{\mathrm{f}}^{\mathrm{insp}} - M_{\mathrm{f}}^{\mathrm{postinsp}}}{M_{\mathrm{f}}^{\mathrm{insp}} + M_{\mathrm{f}}^{\mathrm{postinsp}}}, \tag{4a}$$

$$\frac{\Delta \chi_{\mathrm{f}}}{\bar{\chi}_{\mathrm{f}}} \coloneqq 2\frac{\chi_{\mathrm{f}}^{\mathrm{insp}} - \chi_{\mathrm{f}}^{\mathrm{postinsp}}}{\chi_{\mathrm{f}}^{\mathrm{insp}} + \chi_{\mathrm{f}}^{\mathrm{postinsp}}}, \tag{4b}$$

where the "insp" and "postinsp" superscripts denote the estimates obtained from the inspiral and postinspiral portions of the signal. These fractional deviations should be consistent with zero if the waveform model is a good description of the observed signal. As in [10,11], we present results with a flat prior on $\Delta M_{\mathrm{f}}/\bar{M}_{\mathrm{f}}$ and $\Delta\chi_{\mathrm{f}}/\bar{\chi}_{\mathrm{f}}$.

We now describe how we obtain f_{cut}. As in the applications of the test to real gravitational wave data, e.g., [7,10,11], we use the median values from the analysis of the simulated observation using GR waveform models. Additionally, for comparison, we also apply the test using the same f_{cut} one obtains from the GR simulated observations corresponding to the modified GR cases (which are close to the values one would obtain from the simulated waveforms themselves, as one would expect). We show these comparisons of f_{cut} and the resulting IMR consistency results in the Appendix.

For comparison with the SNRs recovered by the other tests, we compute a combined SNR from the inspiral and postinspiral analyses. To do this, we note that the SNRs from the two portions add in quadrature, so we can compute a probability density for them as is done for the fractional deviations in the Appendix of [84], giving

$$P(\rho_{\mathrm{tot}}) = \int_0^{\rho_{\mathrm{tot}}} \frac{P_{\mathrm{insp}}(\rho_{\mathrm{insp}}) P_{\mathrm{postinsp}}\left(\sqrt{\rho_{\mathrm{tot}}^2 - \rho_{\mathrm{insp}}^2}\right)}{\sqrt{\rho_{\mathrm{tot}}^2 - \rho_{\mathrm{insp}}^2}} \rho_{\mathrm{tot}} d\rho_{\mathrm{insp}}, \tag{5}$$

where $P(\rho_{\mathrm{tot}})$, $P_{\mathrm{insp}}(\rho_{\mathrm{insp}})$, and $P_{\mathrm{postinsp}}(\rho_{\mathrm{postinsp}})$ denote the probability densities for the total SNR and the SNRs in the inspiral and postinspiral, respectively. Specifically, we change variables from $\{\rho_{\mathrm{insp}}, \rho_{\mathrm{postinsp}}\}$ to $\{\rho_{\mathrm{insp}}, \rho_{\mathrm{tot}}\}$, where $\rho_{\mathrm{tot}} = \sqrt{\rho_{\mathrm{insp}}^2 + \rho_{\mathrm{postinsp}}^2}$, and marginalize over ρ_{insp}, noting that it is nonnegative and at most ρ_{tot}.

D. Parametrized tests of gravitational-wave generation

In an alternative theory of gravity, the equations of motion describing the orbital evolution of a coalescing compact binary will in general be different from those in GR. Thus, the frequency evolution of the GW emission will in general be different from the one predicted by GR. The parametrized tests aim to detect GR violations by allowing for deviations in the frequency-domain phase coefficients of the GR waveform models, as initially proposed in [103–106]

The early inspiral dynamics of the compact binary is described with good accuracy using the well known PN approximation to GR (see, e.g., Ref. [107]). The frequency-domain phase in the PN approximation is obtained as an expansion in powers of the velocity parameter v, which is defined as a function of frequency f, i.e., $v = (\pi M_z f)^{1/3}$, where M_z is the total redshifted mass of the binary. In the frequency domain, the phasing for a nonprecessing (i.e., aligned-spin) binary can be schematically written as (omitting additive constants and the effect of a time shift)

[2]As in Ref. [3], we average the fits and augment the aligned-spin final spin fits with the contribution from in-plane spins [102]. However, as in [7,10,11], we do not evolve the spins before applying the fits, for technical reasons.

$$\Phi(f) = \frac{3}{128\eta v^5}\sum_k (\varphi_k v^k + \varphi_{kl} v^k \ln v), \tag{6}$$

where $\eta \coloneqq m_1 m_2/(m_1+m_2)^2$ is the symmetric mass ratio ($m_{1,2}$ are the individual masses of the components of the binary) and the summation is taken over all the PN orders for which we know the phase evolution. The terms φ_k and φ_{kl} are PN coefficients which encode various physical effects in the dynamics of the binary and hence are functions of binary parameters such as masses and spins.

In the IMRPhenomPv2 waveform model, the inspiral portion of the GW phasing is described using the PN phasing augmented by phenomenological coefficients obtained by fitting to numerical relativity waveforms. Similarly, the late inspiral and merger-ringdown portions of the GW phasing are described by phenomenological expressions in powers of frequency, which include the late inspiral (*a.k.a.* intermediate) coefficients β_i and merger-ringdown coefficients α_i. The dependence of α_i and β_i on binary parameters is also obtained by fitting to numerical relativity waveforms. For convenience, we denote the phasing coefficients in all three regimes collectively by p_i.

One version of the parametrized test is the TIGER approach [106,108,109], whose implementation applied to LIGO-Virgo detections to date uses IMRPhenomPv2 as the underlying GR model. In this method, one introduces dimensionless fractional deviations $\delta\hat{p}_k$ in each phasing coefficient p_k as fractional deviations from GR such that $p_k \to p_k^{\rm NS}(1+\delta\hat{p}_k) + p_k^{\rm S}$. Here the superscripts NS and S denote the nonspinning and spinning parts of the GR phasing coefficients. The fractional deviations are only scaled by the nonspinning part of the coefficients to avoid cases where the spin contributions cause the GR coefficient to vanish. Additionally, these coefficients only modify the phasing of the underlying aligned-spin waveform model (IMRPhenomD), not the precessing dynamics used to twist up those waveforms to obtain the final IMRPhenomPv2 precessing waveform model.

A second approach is the FTA approach, introduced in [6], which is not tied to a specific waveform model, but only considers the inspiral PN coefficients. Here it is applied using the SEOBNRv4_ROM waveform model as the GR baseline, as in the LIGO-Virgo analyses [7,10,11]. Unlike the TIGER approach, where the deviations in the early-inspiral and late-inspiral coefficients affect all the higher-frequency portions of the waveform through the C^2 matching used to stitch together the different parts of the IMRPhenomD model [86], in the FTA approach, the deviations are tapered to zero at some frequency. Here, as in the LIGO-Virgo analyses [7,10,11], this frequency is chosen to be 0.35 times the peak frequency of the SEOBNRv4 model. This choice is designed to be consistent with the end of the early-inspiral portion of IMRPhenomD ($GM_z f/c^3 = 0.018$) [86] used in TIGER. The FTA deviations are parametrized in terms of the spinning PN coefficients, unlike TIGER's use of just the nonspinning part. They are then mapped to the TIGER parametrization to obtain the final results, as discussed in [7,10,11]. Since both tests only modify the nonprecessing phasing, their results are then directly comparable.

If the waveform model used (e.g., IMRPhenomPv2 or SEOBNRv4_ROM) is a good description of the signal, all of the fractional deviations introduced in the test should be consistent with zero. Thus, one ideally would constrain all the $\delta\hat{p}_k$ simultaneously, as one would in general expect all of them (at least above some PN order) to simultaneously deviate from their GR values if there is a deviation from GR. However, the fractional deviations are highly correlated, so in practice it is very difficult to measure all of them with current SNRs—see [1] for an explicit illustration of this with GW150914. Nevertheless, recent works have shown that the ability to constrain multiple PN coefficients together can be improved with multiband observations by LISA and third-generation ground-based detectors [110,111], and/or with the use of principal component analysis [112,113]. However, here we follow the procedure in the most recent LIGO-Virgo testing GR papers [7,10,11] and only vary one deviation parameter at a time. This has been shown to be sufficient to detect at least some deviations from GR [109].

For this first study, we consider the 2PN early-inspiral coefficient $\delta\hat{\varphi}_4$, since it corresponds to the leading order of the deviation from GR in the inspiral phasing of our modified EOB waveforms. We also consider the 1PN coefficient $\delta\hat{\varphi}_2$ in a few cases, since it corresponds to the PN order of the massive graviton dephasing we use for simulated observations and the modified dispersion relation test. We also consider the TIGER late-inspiral and merger-ringdown parameters $\delta\hat{\beta}_2$ and $\delta\hat{\alpha}_2$ since they are somewhat better constrained than the other late-inspiral and merger-ringdown parameters, respectively (see the Appendix of [7]).

E. Modified dispersion test

General relativity predicts that GW propagation is nondispersive. That is, the velocity of propagation is independent of the waves' frequency. This property is equivalent to a massless graviton (using quantum terminology here and in the following for convenience, though we are only concerned with classical effects here). On the other hand, certain alternative theories of gravity predict a massive graviton or other dispersion effects as the waves travel from the source to the observer (see, e.g., Refs. [3,114]). We thus consider a parametrized dispersion relation, following [114], which encompasses the leading predictions of a number of different alternative theories,

$$E^2 = p^2c^2 + A_\alpha p^\alpha c^\alpha. \tag{7}$$

Here A_α is the amplitude of the modified dispersion (zero in GR) and has dimension of $[\text{Energy}]^{2-\alpha}$; α is a

dimensionless constant. In particular, $\alpha = 0$ and $A_0 > 0$ corresponds to a massive graviton with mass $m_g = A_0^{1/2}/c^2$. We will frequently use the dimensionless quantity $\tilde{A}_0 \coloneqq A_0/(10^{-44}\ \mathrm{eV}^2)$, since this is a convenient scale for the amplitudes we are considering.

As discussed in Ref. [7], one can reasonably take gravitational waves near the source to be those predicted by GR to a very good approximation. The only modification to the waveform is the dephasing that builds up over the waves' propagation to Earth [see [7] for the explicit expressions, though the exponent in that paper's Eq. (4) should be $1/(2-\alpha)$, as noted in [10]].[3] For instance, in the massive graviton case, the length scale of the Yukawa modification to the Newtonian potential is constrained to be much larger than the size of the binary, so this modification's effect on the binary's dynamics is negligible. For this first analysis, we restrict to the case $\alpha = 0$, thus including the massive graviton case, though we also allow for $A_0 < 0$ along with $A_0 > 0$.

As in [3,7,10,11], we sample separately for $A_0 > 0$ and $A_0 < 0$ (the sampling is carried out in the logarithm of an effective wavelength and the results are then reweighted to a flat prior in A_0). We also combine together the $A_0 > 0$ and $A_0 < 0$ probability distributions (weighted by their evidences) to allow us to quote the quantile of the distribution corresponding to the GR value of $A_0 = 0$, again as in [7,10,11], as well as the SNR distribution.

The dephasing from the modified dispersion relation maps onto a modified PN coefficient in the inspiral for certain values of α, notably corresponding to a modified 1PN coefficient for the $\alpha = 0$ case we consider. However, the modified dispersion dephasing affects the entire waveform, and thus is quite different from the TIGER and FTA modifications to the PN coefficients described in Sec. II D, which are only applied to the inspiral portion of the waveform.

III. SIMULATED OBSERVATIONS

We consider a variety of simulated observations, both with and without deviations from GR. Specifically, we consider waveforms with the modifications used in both the parametrized tests of GW generation and in the modified dispersion tests and their GR analogs, given by the IMRPhenomPv2 waveform model. We still use IMRPhenomPv2 as the base waveform in the FTA case, instead of SEOBNRv4_ROM, which is used in the application of the test, to avoid introducing another GR model. We also consider the EOB waveforms used to check the performance of the IMR consistency test in [83,84], described in more detail below, as well as their GR analogs. The modified EOB waveforms are only available for nonspinning systems, since they are obtained by modifying an EOB code for nonspinning binaries. Thus, we only consider simulated observations of nonspinning binaries in this paper. Additionally, since we use waveform models that only contain the dominant quadrupolar modes of the GW signal to analyze these waveforms, we just include these modes of the modified EOB waveforms, to avoid any systematics from omitted higher modes, though these would be expected to be minor for the nonspinning close-to-equal-mass cases we consider.

A. EOB waveforms with modified energy flux

As detailed in Ref. [84], the modified EOB waveforms modify the energy flux in the IHES EOB model [116,117] by multiplying the (ℓ, m) modes of the waveform that first enter the energy flux at 2PN, viz., the $(3, \pm 2)$, $(4, \pm 4)$, and $(4, \pm 2)$ modes, by a factor $a_2^{1/2}$, so the modes' contribution to the energy flux is multiplied by a_2.[4] We then iteratively adjust the mass and spin of the final black hole used to calculate the quasinormal mode (QNM) frequencies for the ringdown model so that the waveforms satisfy energy and angular momentum balance (using modes through $\ell = 7$, to match the highest ℓ in the tabulated QNM results). We still use this older EOB model instead of a more recent one because it is implemented in Matlab and thus easy to modify and has a ringdown model given purely in terms of Kerr QNMs, without any further fits. Updating this modified EOB waveform construction to more recent EOB models that include spin, such as [88,118] (with extensions to higher modes in [119,120] and precession in [121–123]), will be the subject of future work.

To give an idea of the effect of these large deviations in the energy flux on the binary's dynamics, we consider the mass of the final black hole as a fraction of the total mass M_f/M and the dimensionless spin of the final black hole χ_f. For the $a_2 = 400$ case, these are $M_f/M = 0.86$ and $\chi_f = 0.30$, while for $a_2 = 40$ they are $M_f/M = 0.92$ and $\chi_f = 0.57$. Both of these pairs are outside the region of pairs obtainable in GR, shown in Fig. 6 of [84]. For comparison, the final mass and spin in the GR case obtained using the fit in the IHES EOB code are $M_f/M = 0.95$ and $\chi_f = 0.68$, which agree with those obtained from the self-consistent calculation with the GR value of $a_2 = 1$ to the number of digits shown, only differing from them by one in the next decimal place (i.e., by $\sim 10^{-3}$). For all the other non-GR waveforms considered, the radiated energy and angular momentum are unchanged from their GR values, since only the frequency domain phase is altered, and these quantities just depend on the frequency domain amplitude, as one can see by converting the expressions in, e.g., Ref. [124] to the frequency domain using the

[3]We use the same TT + lowP + lensing + ext cosmological parameters from [115] as in [7,10,11].

[4]We call this factor a_2 instead of α_2 (as in [84]) to avoid confusion with the $\delta\hat{\alpha}_2$ parametrized test parameter.

TABLE I. Parameters of the waveforms considered in this study (most given to three significant digits): M_z and D_L are the binary's redshifted total mass and luminosity distance, while "RA" denotes the right ascension. The mass ratio is 1.22 for the (modified) EOB waveforms and 1 for all others and the polarization angle is 3.9 rad in all cases. Each non-GR parameter corresponds to a separate waveform (in the TIGER/FTA case, two different waveforms).

	GR parameters						non-GR parameters		
Name	M_z ($M_\odot$)	D_L (Mpc)	inclination (rad)	RA (rad)	declination (rad)	GPS time (s)	modified EOB a_2	MDR $\tilde{A}_0$	TIGER/FTA $\delta\hat{\varphi}_4$
GW150914-like (M_{72})	72.2	452	2.83	1.68	−1.27	1126259462	400	5	−13
							40	1	−2
GW170608-like (M_{20})	19.9	364	2.15	3.64	0.89	1180922494	40	10	−2

Parseval-Plancherel identity (i.e., the unitarity of the Fourier transform) and noting that the time derivatives of the strain become multiplication by the frequency in the frequency domain.

B. Parameter choices

For the binary's GR parameters, we consider a case like GW150914 [81] as well as a lower-mass case like GW170608 [82] (as both of these are consistent with being nonspinning).[5] For the GW150914-like cases, we consider both large and moderate GR deviations, while for the GW170608-like cases, we only consider moderate GR deviations (for the massive graviton [MDR] case, this corresponds to a larger value of A_0 than in the GW150914-like case, since the test is less sensitive for this lower-mass system at a somewhat smaller distance). The mass ratio of the (modified) EOB waveforms was chosen to be the same as a numerical relativity simulation, SXS:BBH:310 [126–128], in case it proved necessary to compare with these waveforms, though this did not end up being the case. The other simulated observations are equal mass, due to a bug in their creation. The parameters of all the simulated observations are given in Table I. The other GR parameters were obtained from the sample from the GWTC-1 [78] release [129] that corresponds to the median of the marginalized total mass distribution. There is a slight difference in the right ascension of the GW150914-like case from what this procedure gives, due to a transcription error—its value for the closest sample is 1.59 rad. Additionally, there was a transcription error in the inclination angle, right ascension, and declination for the GW170608-like case and their values were permuted, while they should have been 2.46, 2.15, and 0.50. The right ascension and declination values given in Table I are wrapped so the declination has a magnitude less than $\pi/2$.

We chose the magnitude of the larger GR deviation for the modified EOB waveforms to be the same as the one used in [83] to illustrate that the IMR consistency test could pick up a self-consistent deviation from GR that is significant but passes the binary pulsar tests and is close enough to GR waveforms that it would likely be detected by a matched filtering pipeline using GR waveforms, since it gives a fairly high SNR in such a pipeline. In the inspiral, this deviation corresponds to a parametrized test modification of $\delta\hat{\varphi}_4 \simeq -14$.[6] We thus chose $\delta\hat{\varphi}_4 = -13$ for the parametrized test simulated observations with the larger GR deviation. (We used -13 instead of -14 by mistake, but found that the resulting TIGER and FTA waveforms are already very different from the GR waveforms, so we did not want to use a larger magnitude deviation.) For the massive graviton simulated observation with the larger deviation, we chose a value of A_0 that is well outside of the posterior probability distribution for GW150914 shown in Fig. 8 of [7].

For the smaller GR deviation, we chose a ten times smaller deviation in the modified EOB waveforms (still twice as large as the deviation used to check the IMR consistency test for a population of detections in [84]) and the corresponding parametrized test deviation rounded to the nearest integer, $\delta\hat{\varphi}_4 = -2$ (so $1 + \delta\hat{\varphi}_4$ is about ten times smaller than in the larger case). In the massive graviton case, we chose the A_0 values to be around the 90% bounds for GW150914 and GW170608 given in Table IV of [7]. Even the smallest massive graviton deviation is still much larger than we would expect given the latest bound of 1.27×10^{-23} eV$/c^2$ on the graviton mass m_g from the analysis of confident GWTC-2 binary black hole events [11], which corresponds to $\tilde{A}_0 \leq 1.61 \times 10^{-2}$, since $m_g = A_0^{1/2}/c^2$. We choose these relatively large values of the graviton mass because we are interested in deviations from GR that could potentially be detected with a single event at the moderate SNRs we consider, while the

[5]We expect that the different total masses of the two cases will lead to a difference in the accuracy to which the parameters can be inferred, due to, e.g., the number of cycles in band, as discussed in, e.g., [125], in addition to the difference in accuracy from the different SNR.

[6]We compute this using the TaylorF2 stationary phase approximation expression for the frequency domain phase in terms of the binary's PN binding energy and (modified) energy flux, as in [130].

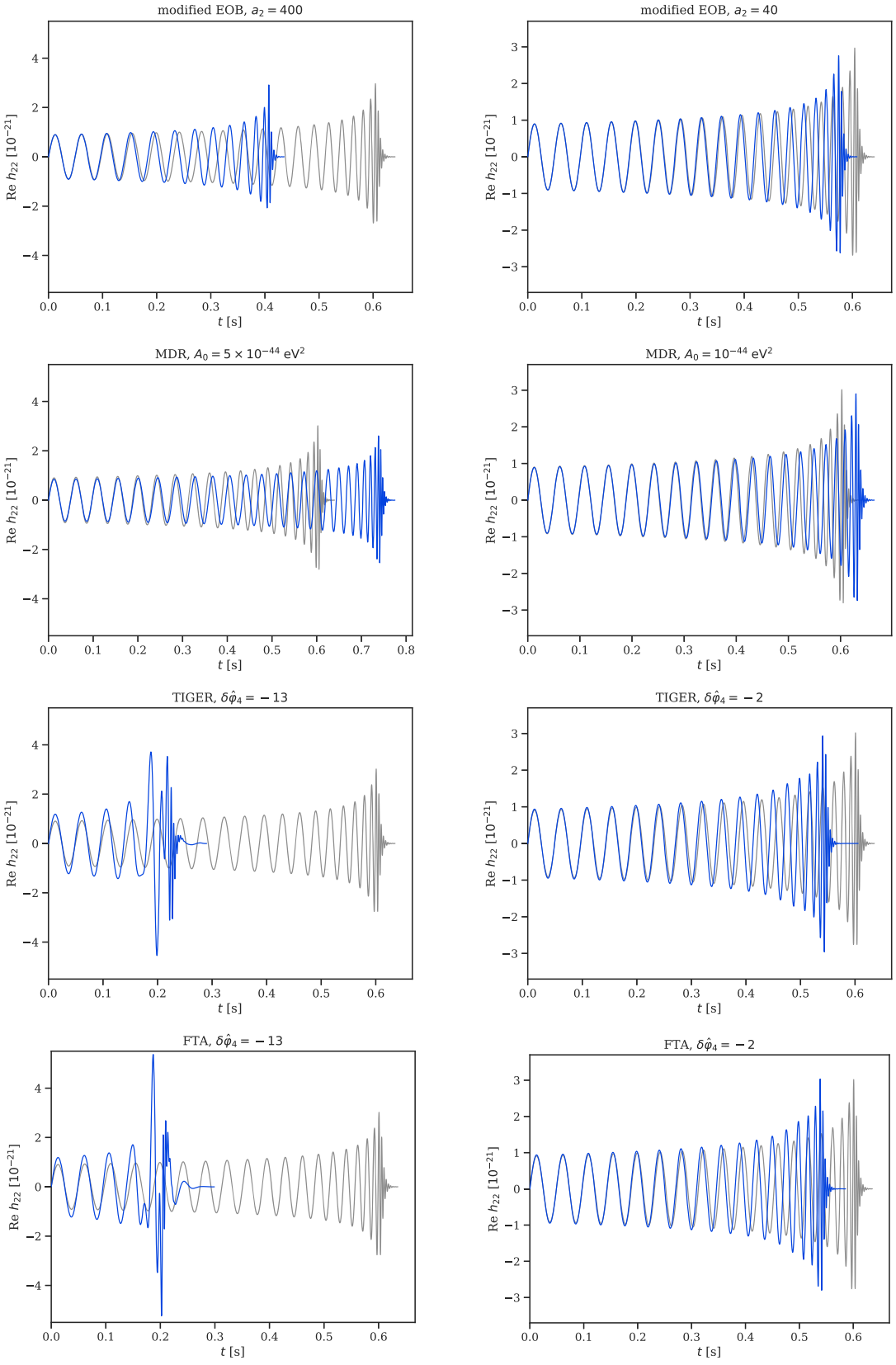

FIG. 1. The GW150914-like waveforms in the time domain. We show the real part of the $\ell = m = 2$ spin-(-2)-weighted spherical harmonic mode of the strain, aligning the non-GR waveforms (blue) with the corresponding GR waveforms (gray) at 20 Hz, which is also the frequency at which the plots start. The larger GR deviation is on the left, and thus those plots have a larger vertical axis range than those on the right.

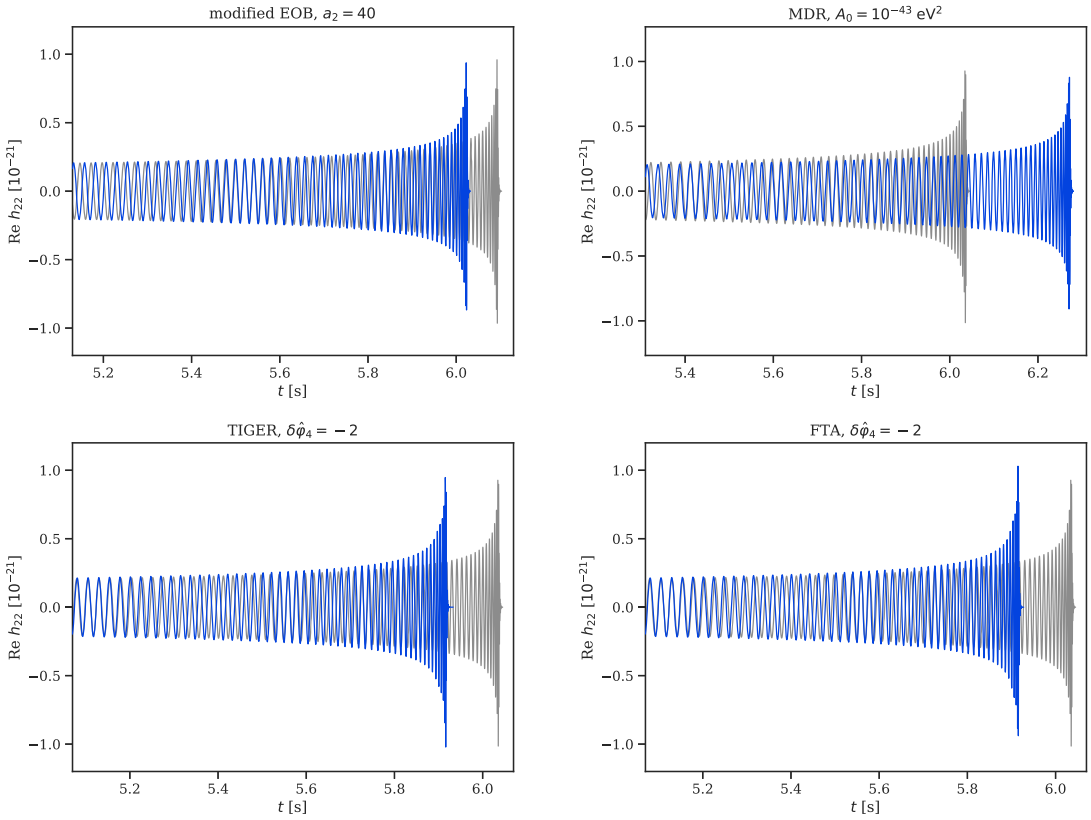

FIG. 2. The analog of Fig. 1 for the GW170608-like waveforms. We still align the waveforms at 20 Hz, but only show the final ∼1 s of the signal.

observational bound on the mass of the graviton is obtained by combining together the constraints from many events.

We plot all these waveforms in the time domain in Figs. 1 and 2 and in the frequency domain in Fig. 3. The frame files for all these simulated observations are available at [131].

IV. RESULTS

We now analyze all the simulated observations from Sec. III with the tests described in Sec. II. In all cases we use a three-detector LIGO-Virgo network with the O3low noise curves from [12] used to construct the likelihood. However, the simulated observations themselves do not contain any noise, so we are effectively averaging over noise realizations [132]. We use a low-frequency cutoff of 20 Hz, so the GW150914-like simulated observations have network optimal signal-to-noise ratios (SNRs) of 54 (53 for the GR EOB simulated observation), except for the modified EOB simulated observation with the larger (smaller) GR deviation, which has a network optimal SNR of 40 (50). (The GR deviations in the non-EOB cases do not affect the SNR, since they do not affect the frequency-domain amplitude.) For the GW170608-like simulated observations, the network optimal SNRs are all 21 (though there are small differences between the Phenom and EOB GR waveforms that are hidden by the rounding to the nearest integer), except for the IHES modified GR simulated observation, which has a network optimal SNR of 20. We use the same priors here as in the LIGO and Virgo collaborations' application of these tests to GW150914 and GW170608 in [7], though we had to increase some prior ranges to account for wider posteriors due to the GR deviations. In particular, we choose flat priors on the IMR consistency test, TIGER, FTA, and MDR deviation parameters.

We give a summary of the results in Figs. 4, 5, and 6 and Table II. The GR quantiles in the table are the quantile at which the GR value of the test is recovered. They are 2D for the IMR consistency test, where smaller values indicate better consistency with GR, and one-dimensional for all other tests, where values around 50% indicate better consistency with GR. We see that the tests all find that the GR simulated observations are consistent with GR within the 90% credible level and discuss the results on the non-GR simulated observations in detail in the following. We show the network matched-filter SNRs recovered by the various analyses in Figs. 7, 8, and 9. In the notation of Sec. II B, the network matched-filter SNR of data $\mathbf{d}$ with waveform model $\mathbf{h}$ is $\rho_{\mathrm{MF}} \coloneqq \langle \mathbf{d}|\mathbf{h}\rangle/\sqrt{\langle \mathbf{h}|\mathbf{h}\rangle}$. Finally, we show the recovery of the final mass and spin (for the

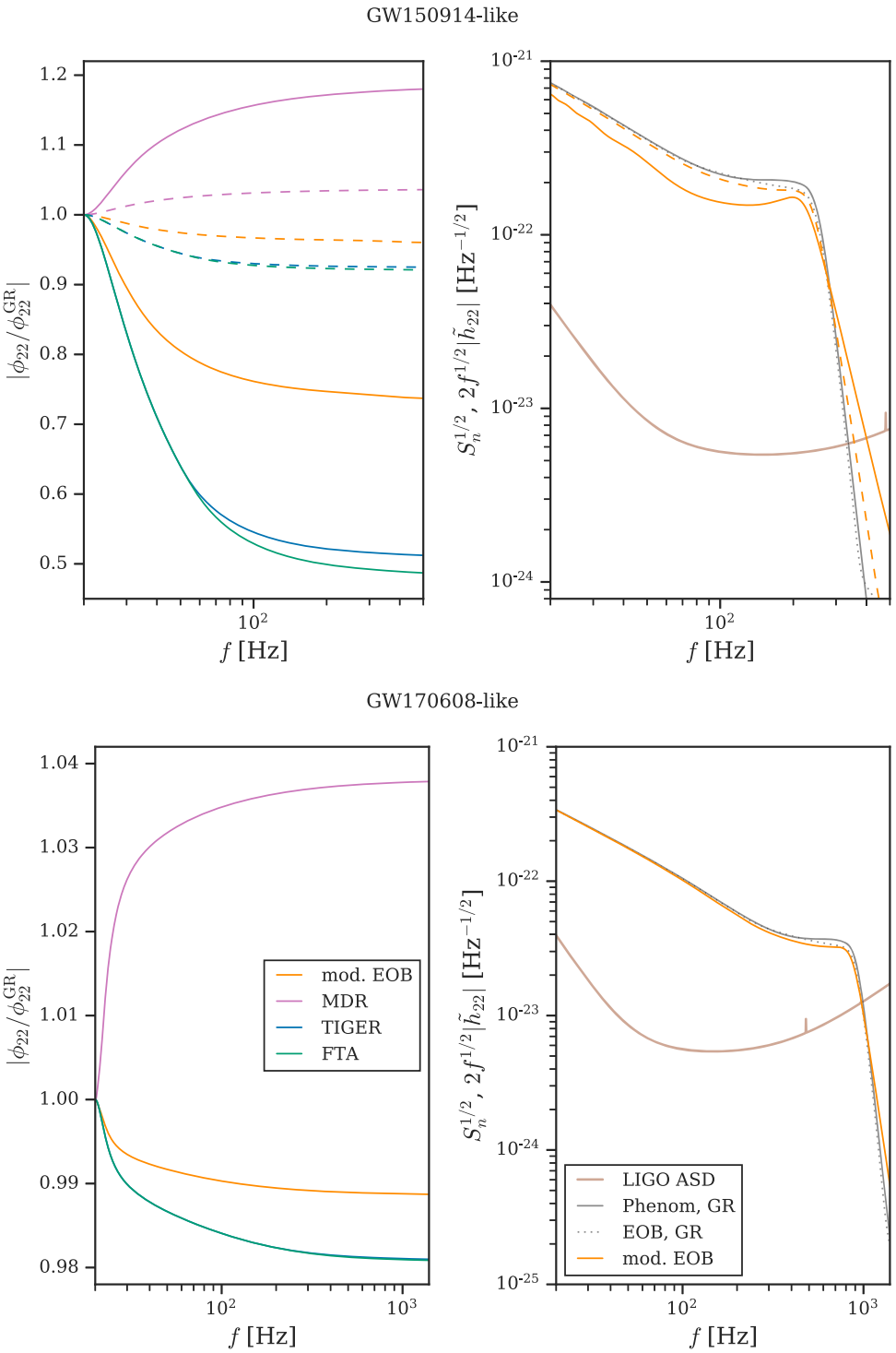

FIG. 3. The GW150914-like and GW170608-like waveforms in the frequency domain. Each of the pairs of plots shows the (frequency-domain) phase (left) and amplitude (right) of the real part of the $\ell = m = 2$ spin-(-2)-weighted spherical harmonic mode of the strain. We align the phases at 20 Hz and set the time shift so that the phase derivative at 20 Hz is 1 s (a somewhat arbitrary choice selected to reduce sharp gradients in the plot near 20 Hz). We then plot the ratio of the phase of the non-GR waveforms to the corresponding GR phase. The dashed lines correspond to the waveforms with smaller GR violations. In the GW170608-like case, the FTA and TIGER phases are almost identical, but the FTA dephasing is slightly greater, as it is in the other cases. For the non-EOB waveforms, we only plot the amplitude of the GR waveform, since these non-GR waveforms only modify the frequency-domain phase. The amplitude is scaled to show the contributions to the SNR integrand, compared to the amplitude spectral density of the noise (cf., e.g., Fig. 1 in [2]); we also show the LIGO noise curve we use in the analysis, for comparison (the O3low one from [12]).

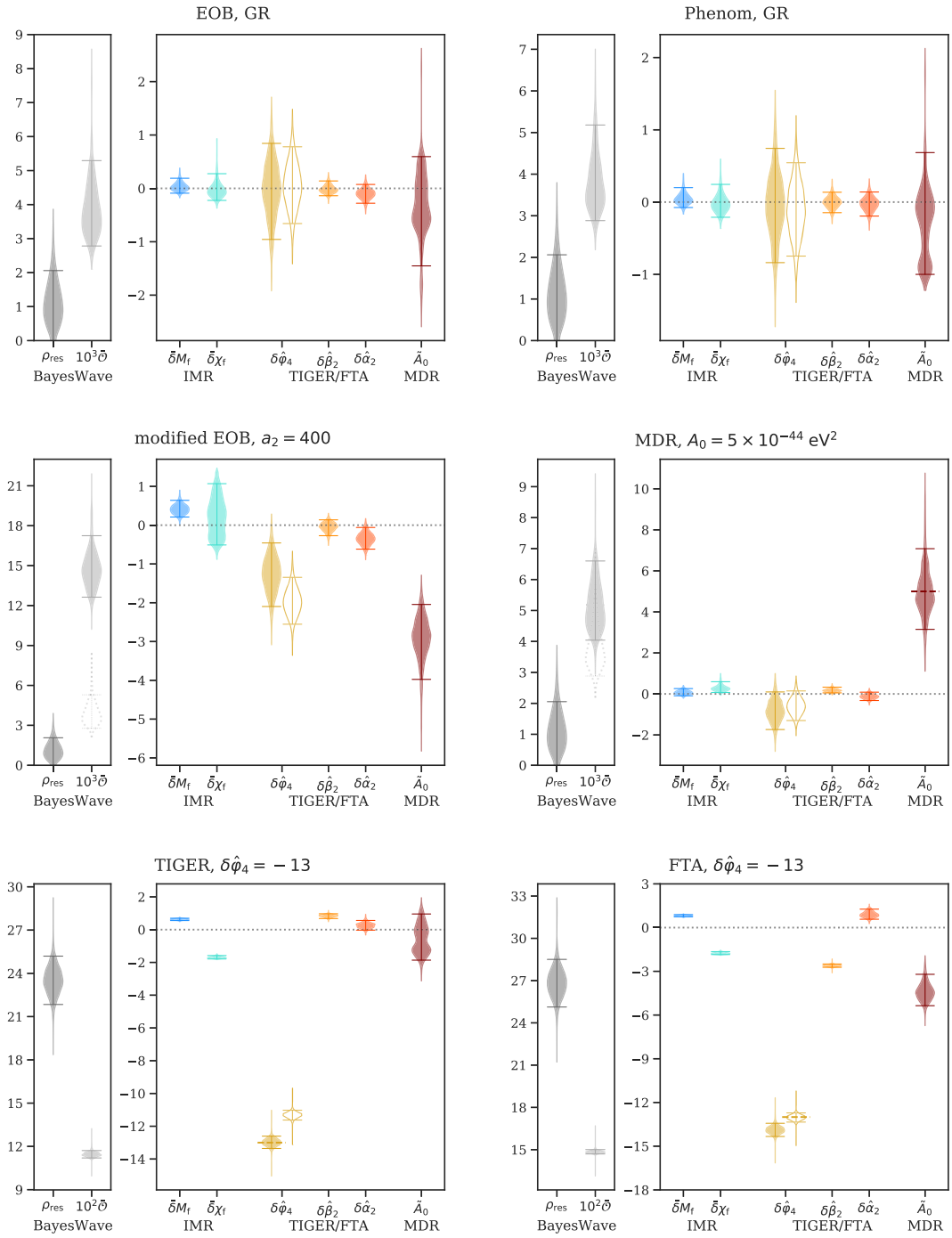

FIG. 4. The results of the various tests on the GW150914-like GR simulated observations and those with the larger GR deviation presented as violin plots of the posteriors on the deviation parameters and the associated 90% credible intervals (CIs). For the residual SNR, we give 90% upper bounds instead of the 90% CI around the median, except for the TIGER and FTA cases where the distribution peaks well away from zero. We write $\bar{\delta}M_{\rm f} \coloneqq \Delta M_{\rm f}/\bar{M}_{\rm f}$ and $\bar{\delta}\chi_{\rm f} \coloneqq \Delta\chi_{\rm f}/\bar{\chi}_{\rm f}$ to save space. We scale $\bar{\mathcal{O}} \coloneqq 1 - \mathcal{O}_{\rm B,L}$ differently for the TIGER and FTA simulated observations, which give much larger values. We mark the GR value of zero with a dotted line for the non-BAYESWAVE tests. Additionally, for the TIGER, FTA, and massive graviton (MDR) simulated observations, we mark the true value of the corresponding test's parameter with a dashed line. For the modified EOB and massive graviton cases, we show the distribution of $\bar{\mathcal{O}}$ for the corresponding GR case as a dotted, unfilled violin, for comparison, since the distributions overlap or are relatively close to doing so.

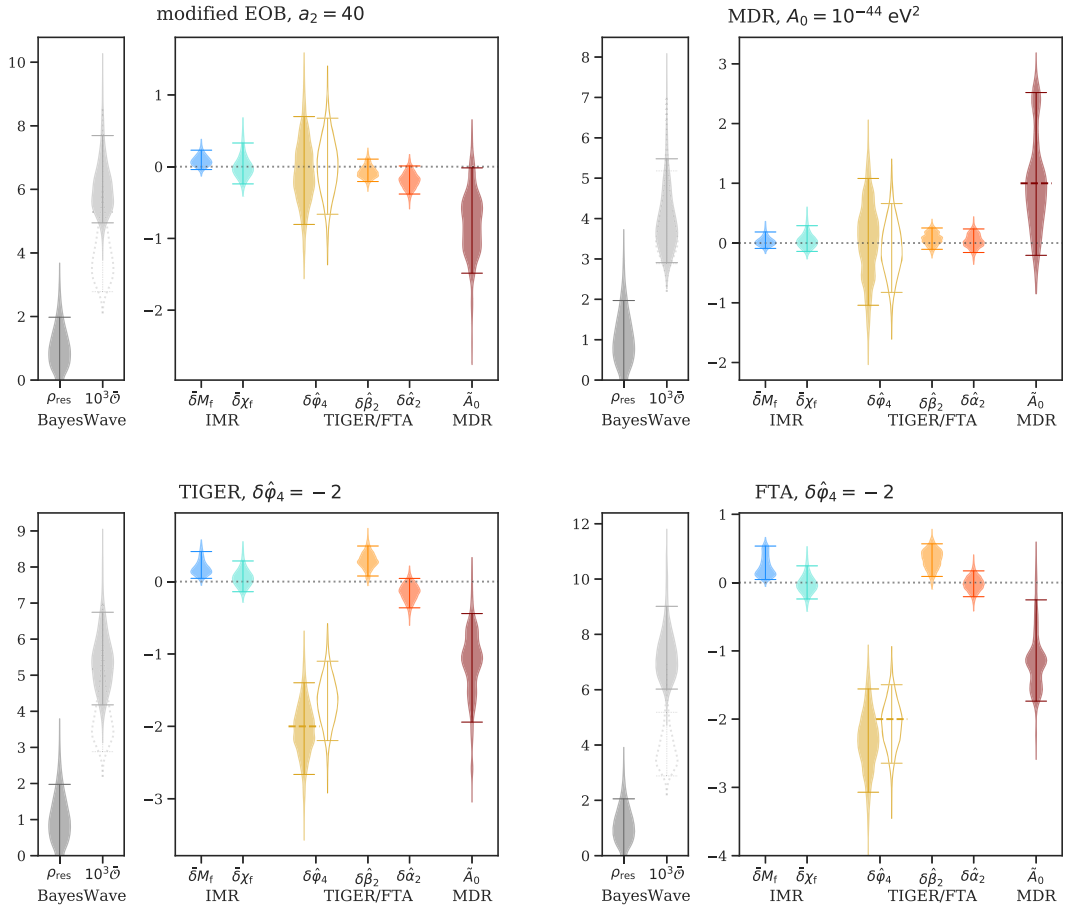

FIG. 5. Like Fig. 4, except for the GW150914-like simulated observations with the smaller GR deviation. Additionally, we are able to scale $\bar{\mathcal{O}}$ the same way for all cases and also show the corresponding GR distribution in all cases.

non-BAYESWAVE tests) in Figs. 10, 11, and 12. These are computed the same way as for the IMR consistency test, as described in Sec. II C.

A. GW150914-like cases

1. Larger GR deviations

We find that all of the GW150914-like simulated observations with the larger GR deviation are identified as not consistent with GR at the 90% credible level by at least three tests (see Table II). For the TIGER and FTA simulated observations, with their very large GR deviations, all the tests pick up the deviations, except surprisingly for the MDR test for the TIGER simulated observation, though the α_2 TIGER case recovers GR just inside the 90% CI. For the modified EOB simulated observation, all the tests recover a strong GR deviation except for the residuals and TIGER β_2 tests. The TIGER and FTA φ_4 tests exclude GR at the 99% credible level or higher, but recover a value of $\delta\hat{\varphi}_4$ that is much smaller than the true value of ~ -14. This is not surprising: These tests are only varying a single PN coefficient, while the simulated observation also has all 3PN and higher coefficients modified and additionally modifies the merger and ringdown. This illustrates that these tests are not designed to measure the true PN coefficients, just to detect deviations from GR. The massive graviton (MDR) simulated observation is identified as a GR violation at the 90% credible level only with the IMR consistency test, TIGER β_2, and MDR analyses.

Since we find that the MDR test finds a strong deviation from GR for the modified EOB simulated observation, we also apply the TIGER and FTA φ_2 analyses to this simulated observation and the massive graviton simulated observation, since this PN coefficient matches the $\alpha = 0$ MDR dephasing in the inspiral. We also ran the TIGER and FTA φ_2 analyses on the GR simulated observations, for comparison. We compare with the TIGER and FTA φ_4 analyses in Fig. 13 and give the analog of Table II in Table III. We find that the φ_2 analyses indeed pick up these GR deviations somewhat more strongly than the φ_4 analyses, except for the TIGER analysis of the massive graviton simulated observation, where the GR quantile is the same as the φ_4 analysis. The TIGER and FTA φ_2 analyses of the modified EOB simulated observation find increased SNR, illustrated in Fig. 14. The SNR

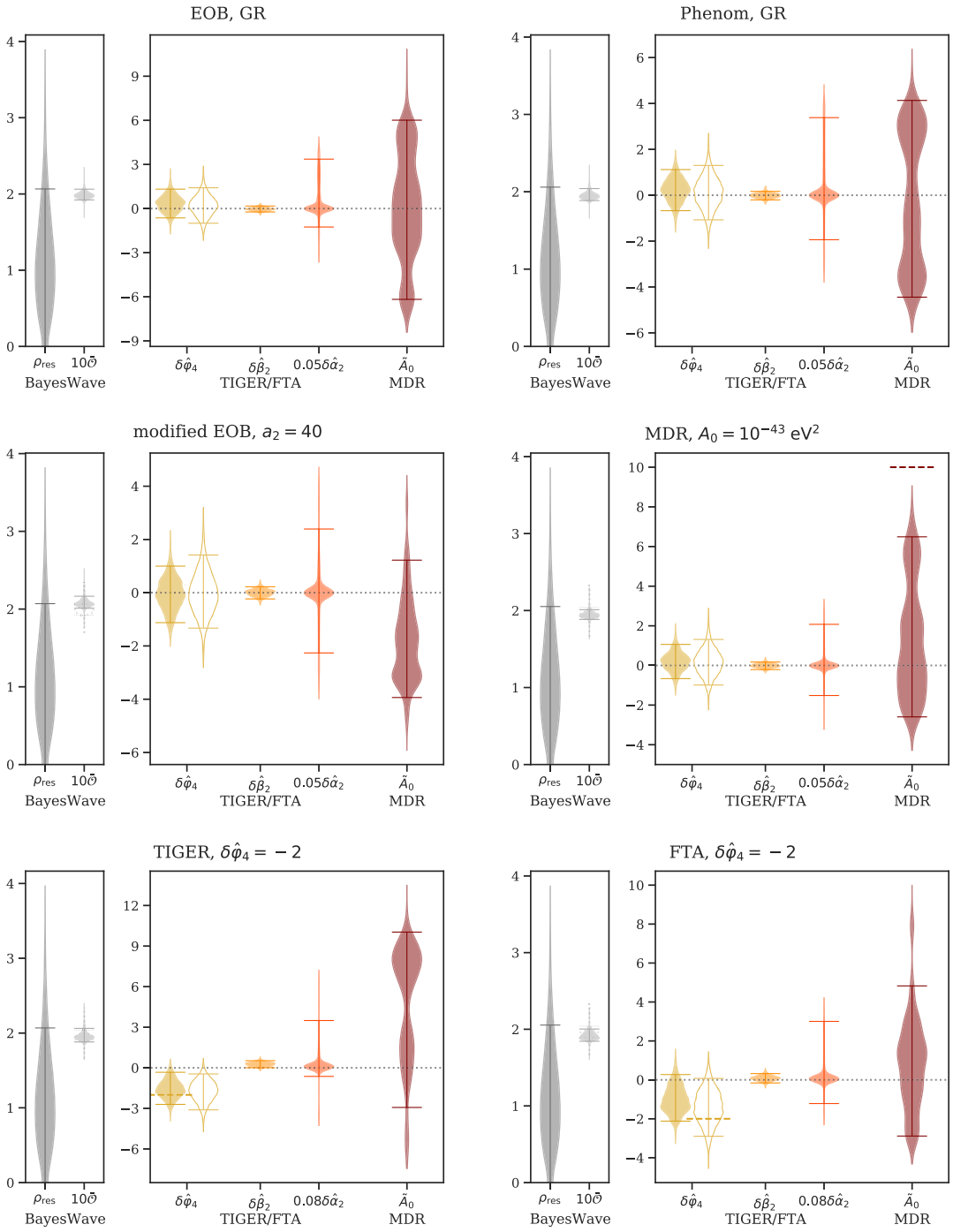

FIG. 6. Like Fig. 4, except for the GW170608-like simulated observations (for which the IMR consistency test is not applicable). Here we scale the broad $\delta\hat{\alpha}_2$ posteriors to make the plot easier to read and scale $\bar{\mathcal{O}}$ by 10 instead of the scalings of 10^2 or 10^3 used in the previous two plots. We also show the corresponding GR distribution of $\bar{\mathcal{O}}$ in all non-GR cases, though there is almost complete overlap except in the modified EOB case.

distributions for the φ_2 analyses of the massive graviton and GR simulated observations are almost identical to those for the φ_4 analyses. For the massive graviton case, the TIGER and FTA φ_2 analyses recover a posterior that excludes the true value of $\delta\hat{\varphi}_2 = -0.66$, with a 90% lower bound of about half the true value, similar to (though less dramatic than) their significant underestimate of the true value of the testing parameter in the modified EOB case.

We find that the residuals test is only able to identify the TIGER and FTA simulated observations, with their very

TABLE II. Summary statistics for the test parameters for all tests on all simulated observations considered. (The IMR consistency test is not applicable to the GW170608-like cases.) For most tests, these are the GR quantiles as well as the median and surrounding 90% CI. For the residual SNR, we give 90% upper bounds instead, except for the TIGER and FTA simulated observations with larger GR deviations, where the probability distribution peaks well away from zero and we give the median and 90% CI. The two BAYESWAVE analyses do not provide GR quantiles. The GW150914-like and GW170608-like simulated observations are denoted by their abbreviations M_{72} and M_{20}, respectively. For the GW150914-like simulated observations, > denotes the case with the larger GR deviation and < the one with the smaller deviation. The GR quantiles are one-dimensional (denoted $Q_{\rm GR}$), so values around 50% indicate good agreement with GR, in all cases except for the IMR consistency test, where they are 2D, denoted $\mathcal{Q}_{\rm GR}$, and values around 0 indicate good agreement with GR. In all cases they are rounded to the nearest percent. We discuss the strength to which GR is excluded when the GR quantile is in the tails of the distribution in the text. We bold the GR quantiles where GR is excluded at the 90% credible level, so where the GR quantile is outside the 90% CI around the median ([5,95]%) for the one-dimensional quantiles.

		Residuals	Reconstructions	IMR consistency			TIGER/FTA								MDR	
							φ_4, TIGER		φ_4, FTA		β_2		α_2			
Simulated observation		$\rho_{\rm res}$	$1-\mathcal{O}_{\rm B,L}$ (10^{-3})	$\mathcal{Q}_{\rm GR}$ (%)	$\frac{\Delta M_{\rm f}}{\overline{M}_{\rm f}}$	$\frac{\Delta\chi_{\rm f}}{\overline{\chi}_{\rm f}}$	$Q_{\rm GR}$ (%)	$\delta\hat{\varphi}_4$	$Q_{\rm GR}$ (%)	$\delta\hat{\varphi}_4$	$Q_{\rm GR}$ (%)	$\delta\hat{\beta}_2$	$Q_{\rm GR}$ (%)	$\delta\hat{\alpha}_2$	$Q_{\rm GR}$ (%)	$\tilde{A}_0$
M_{72}	EOB, GR	<2.1	$3.8^{+1.5}_{-1.0}$	4	$0.0^{+0.2}_{-0.1}$	$0.0^{+0.3}_{-0.2}$	50	$0.0^{+0.8}_{-1.0}$	45	$0.1^{+0.7}_{-0.8}$	62	$0.0^{+0.1}_{-0.1}$	85	$-0.1^{+0.2}_{-0.2}$	69	$-0.3^{+0.9}_{-1.1}$
	Phenom, GR	<2.1	$3.7^{+1.5}_{-0.8}$	3	$0.0^{+0.2}_{-0.1}$	$0.0^{+0.3}_{-0.2}$	54	$0.0^{+0.7}_{-0.8}$	61	$-0.1^{+0.6}_{-0.6}$	52	$0.0^{+0.1}_{-0.1}$	60	$0.0^{+0.1}_{-0.2}$	68	$-0.2^{+0.9}_{-0.8}$
	modified EOB, >	<2.1	$14.6^{+2.6}_{-2.0}$	**100**	$0.4^{+0.2}_{-0.2}$	$0.2^{+0.9}_{-0.7}$	**99**	$-1.3^{+0.8}_{-0.8}$	**100**	$-2.0^{+0.7}_{-0.6}$	64	$0.0^{+0.1}_{-0.3}$	**98**	$-0.3^{+0.2}_{-0.3}$	**100**	$-2.9^{+0.8}_{-1.1}$
	modified EOB, <	<2.0	$6.0^{+1.7}_{-1.0}$	14	$0.1^{+0.1}_{-0.1}$	$0.0^{+0.4}_{-0.2}$	57	$-0.1^{+0.8}_{-0.7}$	50	$0.0^{+0.7}_{-0.7}$	79	$-0.1^{+0.2}_{-0.1}$	94	$-0.2^{+0.2}_{-0.2}$	**95**	$-0.8^{+0.8}_{-0.7}$
	MDR, >	<2.1	$5.1^{+1.5}_{-1.0}$	**90**	$0.1^{+0.2}_{-0.2}$	$0.3^{+0.3}_{-0.2}$	93	$-0.8^{+0.9}_{-0.9}$	91	$-0.6^{+0.8}_{-0.7}$	**1**	$0.2^{+0.1}_{-0.1}$	85	$-0.1^{+0.2}_{-0.2}$	**0**	$4.9^{+2.2}_{-1.7}$
	MDR, <	<2.0	$3.9^{+1.6}_{-0.9}$	15	$0.0^{+0.2}_{-0.1}$	$0.0^{+0.3}_{-0.2}$	45	$0.1^{+1.0}_{-1.1}$	59	$-0.1^{+0.8}_{-0.7}$	25	$0.1^{+0.2}_{-0.2}$	41	$0.0^{+0.2}_{-0.2}$	10	$0.9^{+1.6}_{-1.1}$
	TIGER, >	$23.5^{+1.7}_{-1.7}$	$114.4^{+2.7}_{-2.4}$	**100**	$0.6^{+0.1}_{-0.1}$	$-1.7^{+0.1}_{-0.1}$	**100**	$-13.0^{+0.4}_{-0.4}$	**100**	$-11.3^{+0.3}_{-0.3}$	**0**	$0.8^{+0.2}_{-0.1}$	7	$0.3^{+0.3}_{-0.3}$	71	$-0.6^{+1.6}_{-1.2}$
	TIGER, <	<2.0	$5.3^{+1.4}_{-1.2}$	**98**	$0.2^{+0.2}_{-0.1}$	$0.0^{+0.2}_{-0.2}$	**100**	$-2.0^{+0.6}_{-0.7}$	**100**	$-1.6^{+0.5}_{-0.6}$	**1**	$0.3^{+0.2}_{-0.2}$	89	$-0.1^{+0.1}_{-0.3}$	**99**	$-1.1^{+0.6}_{-0.9}$
	FTA, >	$26.8^{+1.7}_{-1.7}$	$148.3^{+1.7}_{-1.4}$	**100**	$0.8^{+0.1}_{-0.1}$	$-1.8^{+0.1}_{-0.1}$	**100**	$-13.9^{+0.5}_{-0.4}$	**100**	$-13.0^{+0.3}_{-0.3}$	**100**	$-2.6^{+0.1}_{-0.1}$	**0**	$0.9^{+0.4}_{-0.3}$	**100**	$-4.4^{+1.2}_{-0.9}$
	FTA, <	<2.1	$7.2^{+1.8}_{-1.2}$	**97**	$0.2^{+0.3}_{-0.2}$	$0.0^{+0.3}_{-0.2}$	**100**	$-2.3^{+0.7}_{-0.8}$	**100**	$-2.1^{+0.6}_{-0.5}$	**1**	$0.4^{+0.2}_{-0.3}$	56	$0.0^{+0.2}_{-0.2}$	**97**	$-1.2^{+0.9}_{-0.6}$
M_{20}	EOB, GR	<2.1	198^{+8}_{-6}	···	···	···	30	$0.3^{+1.0}_{-0.9}$	38	$0.2^{+1.2}_{-1.2}$	63	$0.0^{+0.2}_{-0.2}$	38	2^{+65}_{-27}	49	$0.1^{+5.9}_{-6.3}$
	Phenom, GR	<2.1	195^{+10}_{-6}	···	···	···	32	$0.2^{+0.9}_{-0.9}$	43	$0.1^{+1.2}_{-1.2}$	56	$0.0^{+0.2}_{-0.2}$	44	1^{+67}_{-40}	53	$-0.4^{+4.5}_{-4.1}$
	modified EOB	<2.1	207^{+10}_{-6}	···	···	···	53	$-0.1^{+1.1}_{-1.0}$	52	$0.0^{+1.4}_{-1.3}$	49	$0.0^{+0.2}_{-0.2}$	49	0^{+48}_{-45}	86	$-2.0^{+3.2}_{-2.0}$
	MDR	<2.1	194^{+7}_{-5}	···	···	···	34	$0.2^{+0.9}_{-0.9}$	41	$0.1^{+1.2}_{-1.1}$	54	$0.0^{+0.2}_{-0.2}$	49	0^{+41}_{-30}	39	$0.9^{+5.6}_{-3.5}$
	TIGER	<2.1	195^{+12}_{-7}	···	···	···	**98**	$-1.6^{+1.3}_{-1.1}$	**99**	$-1.7^{+1.3}_{-1.4}$	**5**	$0.3^{+0.2}_{-0.3}$	33	2^{+42}_{-10}	15	$6.4^{+3.6}_{-9.3}$
	FTA	<2.1	191^{+10}_{-6}	···	···	···	90	$-1.1^{+1.4}_{-1.0}$	94	$-1.4^{+1.5}_{-1.5}$	28	$0.1^{+0.2}_{-0.3}$	35	1^{+37}_{-16}	37	$0.9^{+4.0}_{-3.7}$

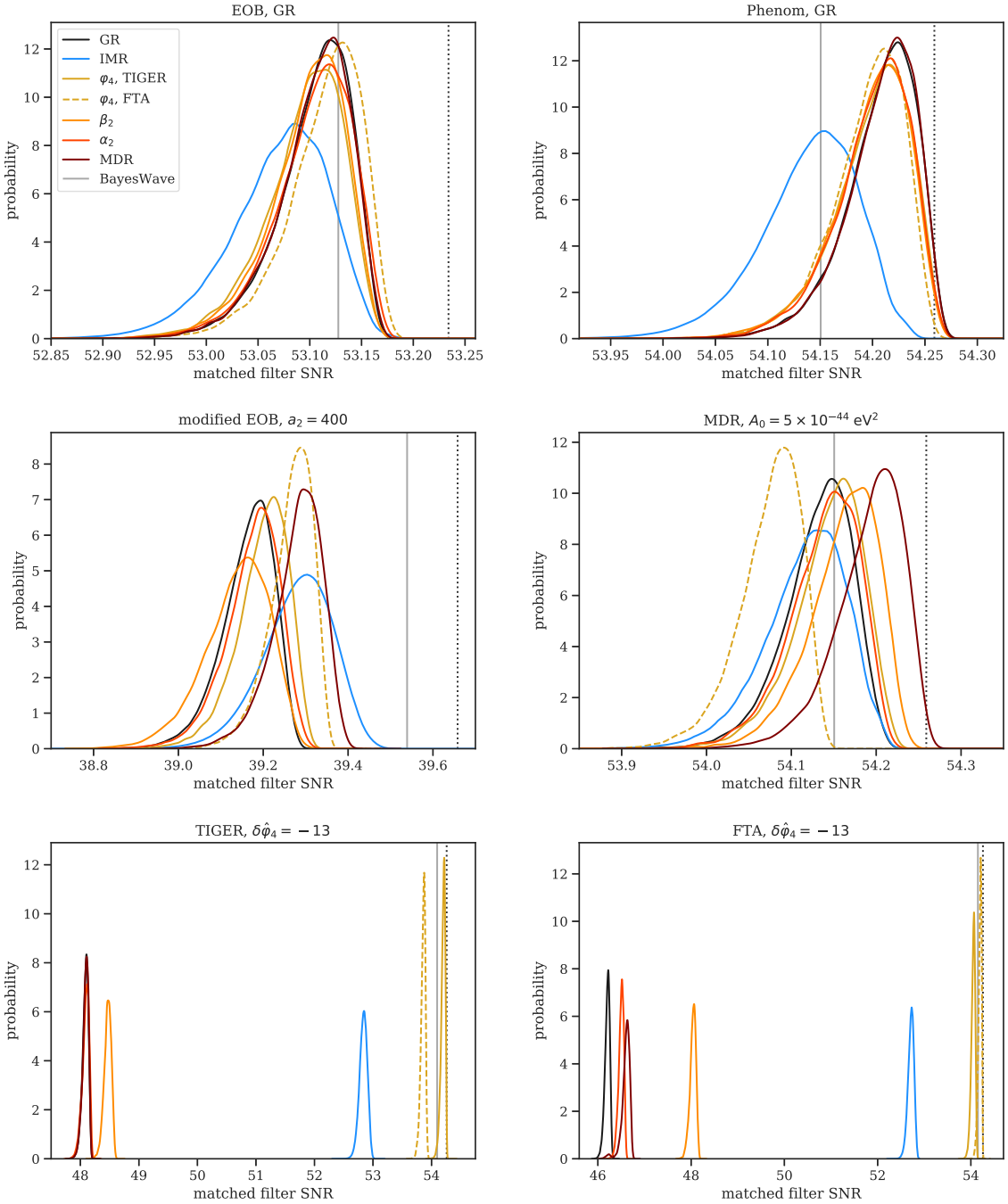

FIG. 7. The posterior distributions of the recovered matched filter SNRs for the various tests applied to the GW150914-like GR simulated observations and those with the larger GR deviation, as well as the matched filter SNR recovered by the median BAYESWAVE reconstruction for these cases. We also show the optimal SNR of the simulated observation, plotted as a vertical dotted line. The IMR results combine together the inspiral and postinspiral posteriors to give a posterior on the SNR for the full frequency range, while the MDR results combine together the positive and negative A_0 results.

large GR violations, as deviations from GR. Even though the modified EOB and massive graviton simulated observations also have significant GR violations that are easily picked up by some of the other tests, the distribution of residual SNRs is almost identical to that for the GR simulated observations. This is in agreement with the results in Fig. 7, which show that the GR analysis is able to recover most of the SNR in those cases. We illustrate the residuals and their BAYESWAVE recovery in a few cases in the bottom panels of Fig. 15, which show the residual detector data in the LIGO Livingston detector and the recovered 90% CI with BAYESWAVE. This illustrates that while BAYESWAVE is able to recover the residual signal very well when it is relatively significant, as for the FTA case, it

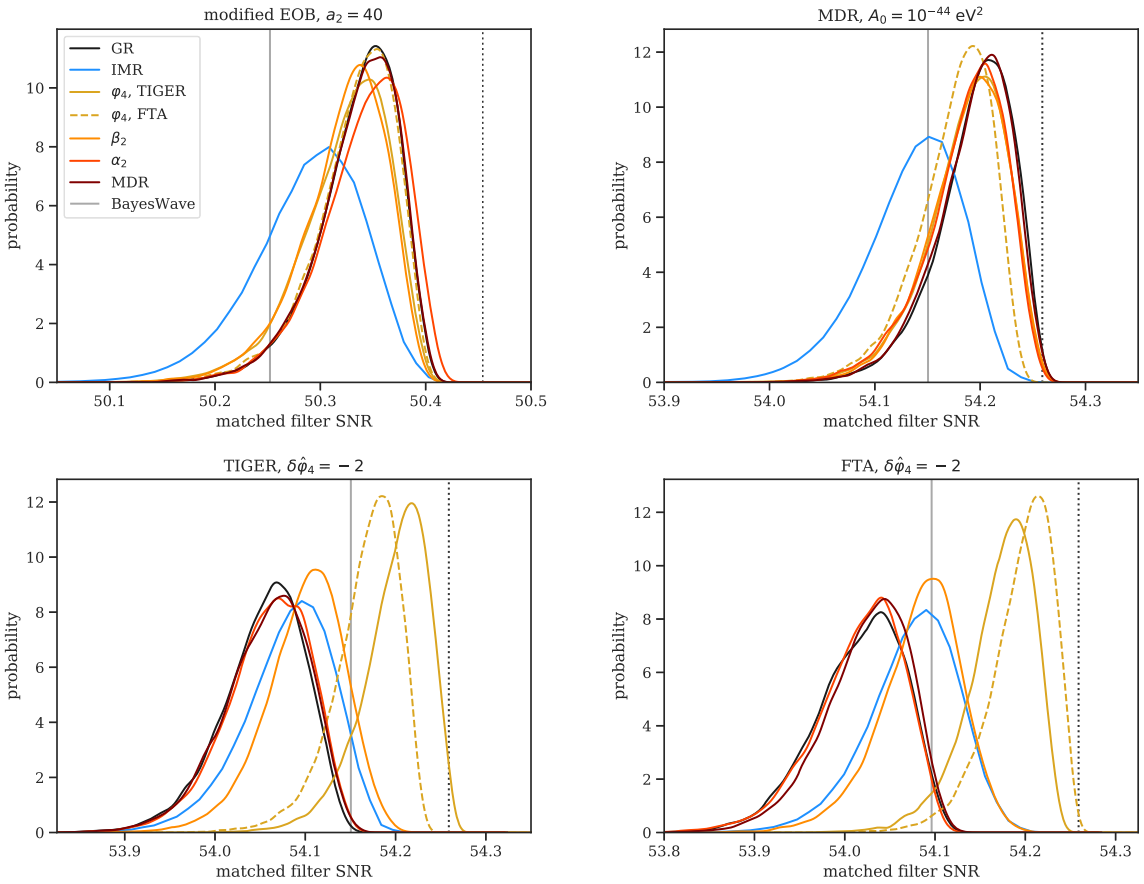

FIG. 8. The analog of Fig. 7 for the GW150914-like simulated observations with the smaller GR deviation.

does not find any coherent signal in the residual for the GR and modified EOB cases with their quite small and relatively small residuals, respectively.

The comparison of reconstructions is more sensitive, finding a distribution of overlaps that is disjoint from the one for the GR simulated observations for the modified EOB case, and a clear shift to larger mismatches for the massive graviton case, as well as clearly picking up the very large TIGER and FTA GR violations. This can be seen qualitatively in the top panels of Fig. 15, where the difference between the BAYESWAVE and LALInference reconstructions increases from left to right with increasing size of the GR deviation, and quantitatively through the overlaps given in Fig. 4 and Table II. The reconstruction is also able to recover the most SNR of any analysis for the modified EOB case, and is second only to the TIGER (FTA) analysis for the TIGER (FTA) simulated observation, as shown in Fig. 7. However, it only recovers about as much SNR as the GR analysis for the massive graviton simulated observation, likely because the dispersion spreads out the waveform, and the reconstructions do better at recovering short waveforms—the modified EOB, TIGER, and FTA waveforms are all significantly shorter than their GR counterparts, as shown in Fig. 1.

For the cases where the GR quantile rounds to 0 or 100% in Tables II and III, it is interesting to consider how strongly GR is excluded. Here we use the scale of Gaussian standard deviations σ, and quote a lower bound of 7σ for cases where the GR quantile is even closer to 0 or 100%. We impose such a lower bound because we have neglected the uncertainties in determining such high credible levels with a finite number of posterior samples, given that these are rather extreme scenarios, so these results should just be taken as roughly indicative of the constraining power of the tests in such cases. For the modified EOB simulated observation, we find that GR is excluded at greater than 7σ by the IMR consistency test, slightly greater than 3σ by the TIGER φ_2 test, by slightly greater than 4.5σ and 4σ in the FTA φ_4 and φ_2 tests, and by slightly greater than 5σ by the MDR test. For the massive graviton simulated observation, GR is excluded at slightly greater than 4σ by the MDR test. For the TIGER and FTA simulated observations, GR is excluded at greater than 7σ by the IMR consistency test as well as the TIGER and FTA φ_4 tests and the TIGER β_2 test. For the FTA simulated observation, GR is also excluded at greater than 7σ by the TIGER α_2 test and at slightly greater than 4σ by the MDR test.

We now consider the recovery of the GR parameters. We start by considering the TIGER simulated observation,

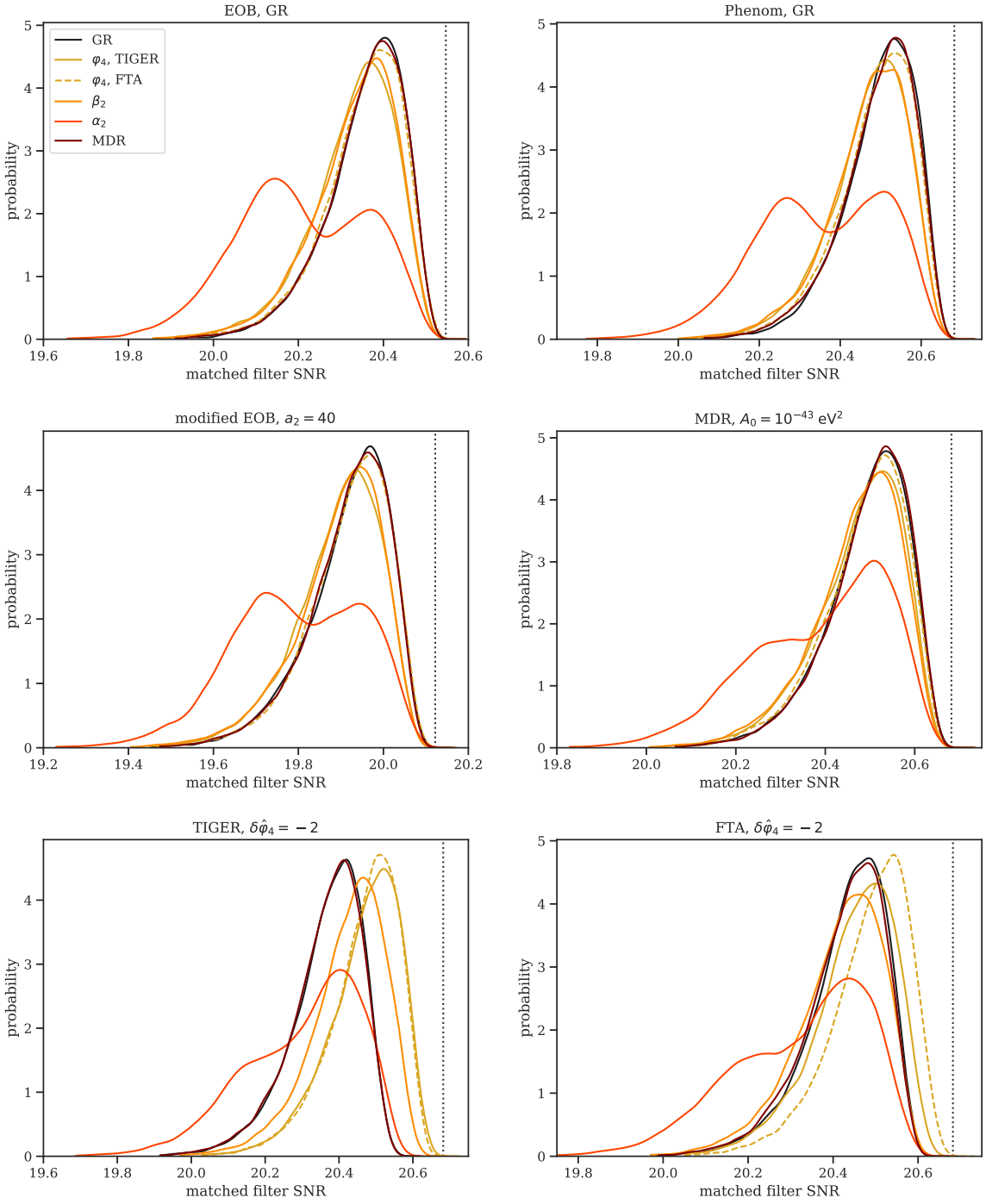

FIG. 9. The analog of Fig. 7 for the GW170608-like simulated observations. Here we do not show the SNR recovered by the median BAYESWAVE reconstruction since it is much smaller than the SNRs plotted for these spread-out signals, as can be seen from the overlaps plotted in Fig. 6 and given in Table II.

where the MDR analysis does not find a GR deviation, recovering an unequal-mass (mass ratio of $0.66^{+0.08}_{-0.08}$, giving the median and surrounding 90% CI) precessing system with a nearly edge-on inclination and a highly spinning primary (the primary spin posterior rails strongly against the high-spin prior bound) with most of the primary spin in the orbital plane. Both signs of A_0 give very similar posteriors, which are also very similar to those from the GR recovery. For instance, the mass ratio median and 90% CI is the same for the MDR recovery with both signs of A_0 and the GR recovery to the precision quoted above. There is also a much larger difference in the recovered SNRs for the different tests for the TIGER and FTA simulated observations than for the others—see Fig. 7. This figure illustrates

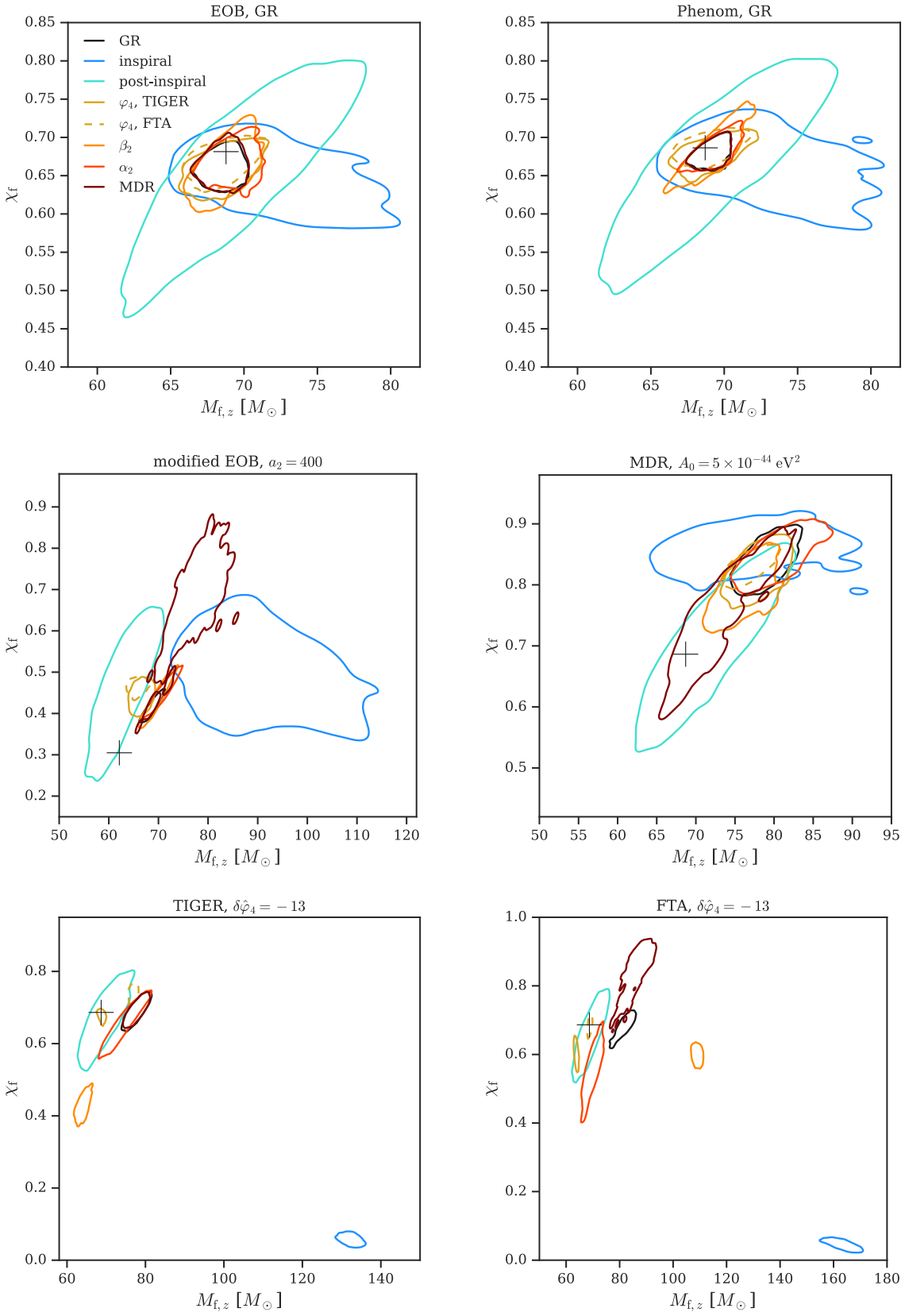

FIG. 10. The 90% credible regions of the joint posterior distributions of the recovered (redshifted) final mass and spin for the GW150914-like GR and larger GR deviation simulated observations, along with the values of the simulated observations, plotted as plus signs.

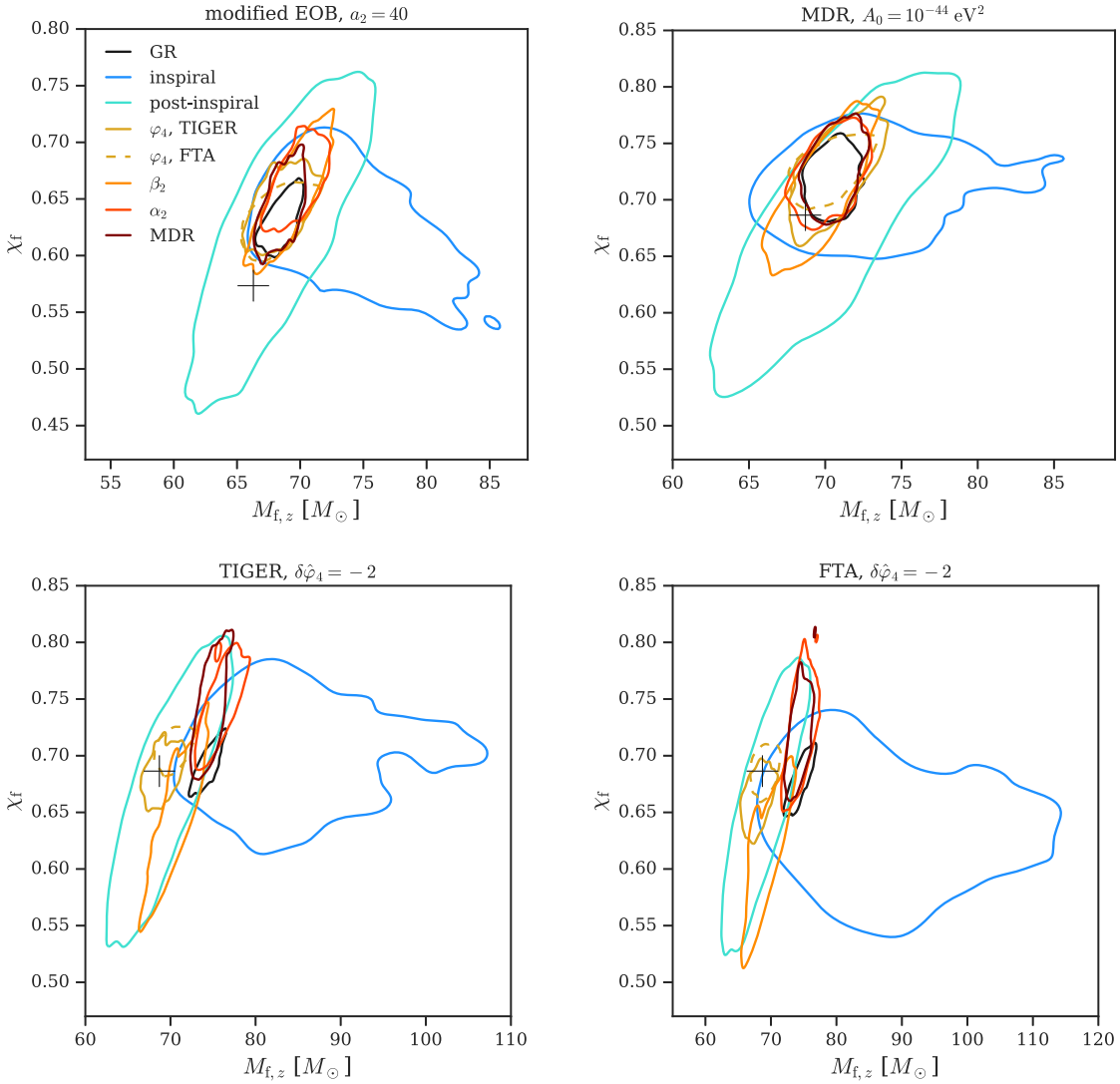

FIG. 11. The analog of Fig. 10 for the GW150914-like smaller GR deviation simulated observations.

that the MDR recovery of the TIGER simulated observation finds a matched filter SNR very similar to the GR recovery, as would be expected, given the similarity of the posteriors for other parameters.

Now considering the modified EOB simulated observation, we find that all the tests except the IMR consistency test inspiral and MDR $A_0 < 0$ cases find close to equal masses, with an inclination angle and distance close to the true ones, but favor large antialigned spins to give a large negative effective spin[7] and thus reduce the length of the signal and the final spin from their nonspinning GR values. The IMR consistency test inspiral recovery favors an unequal-mass system (mass ratio posterior peaking around 0.3) with a small spin on the larger black hole and the smaller black hole's spin unconstrained. The MDR $A_0 < 0$ case finds large in-plane spins with an effective spin posterior that peaks close to zero. In both cases, they favor a slightly smaller inclination angle and larger distance than the true values. These two tests also recover slightly more matched filter SNR than the other tests—see Fig. 7.

We also consider how well the (redshifted) final mass and spin are recovered by the non-BAYESWAVE analyses, plotting the values of the simulated observations and the joint posterior distributions in Fig. 10. We find that the MDR, TIGER, and FTA analyses always recover the true value in the 90% credible region for their associated simulated observations, as does the IMR consistency test postinspiral analysis in all cases (albeit just barely for the

[7]The effective spin is defined by $\chi_{\rm eff} \coloneqq (m_1\chi_1^{\parallel} + m_2\chi_2^{\parallel})/(m_1 + m_2)$, where m_A and $\chi_A^{\parallel}$ ($A \in \{1, 2\}$) are the holes' masses and components of the dimensionless spins parallel to the (Newtonian) orbital angular momentum, respectively. We consider this quantity here since it is a simple, well-measured combination of the spins that is closely related to the dominant spin-orbit coupling (see, e.g., Refs. [133,134]).

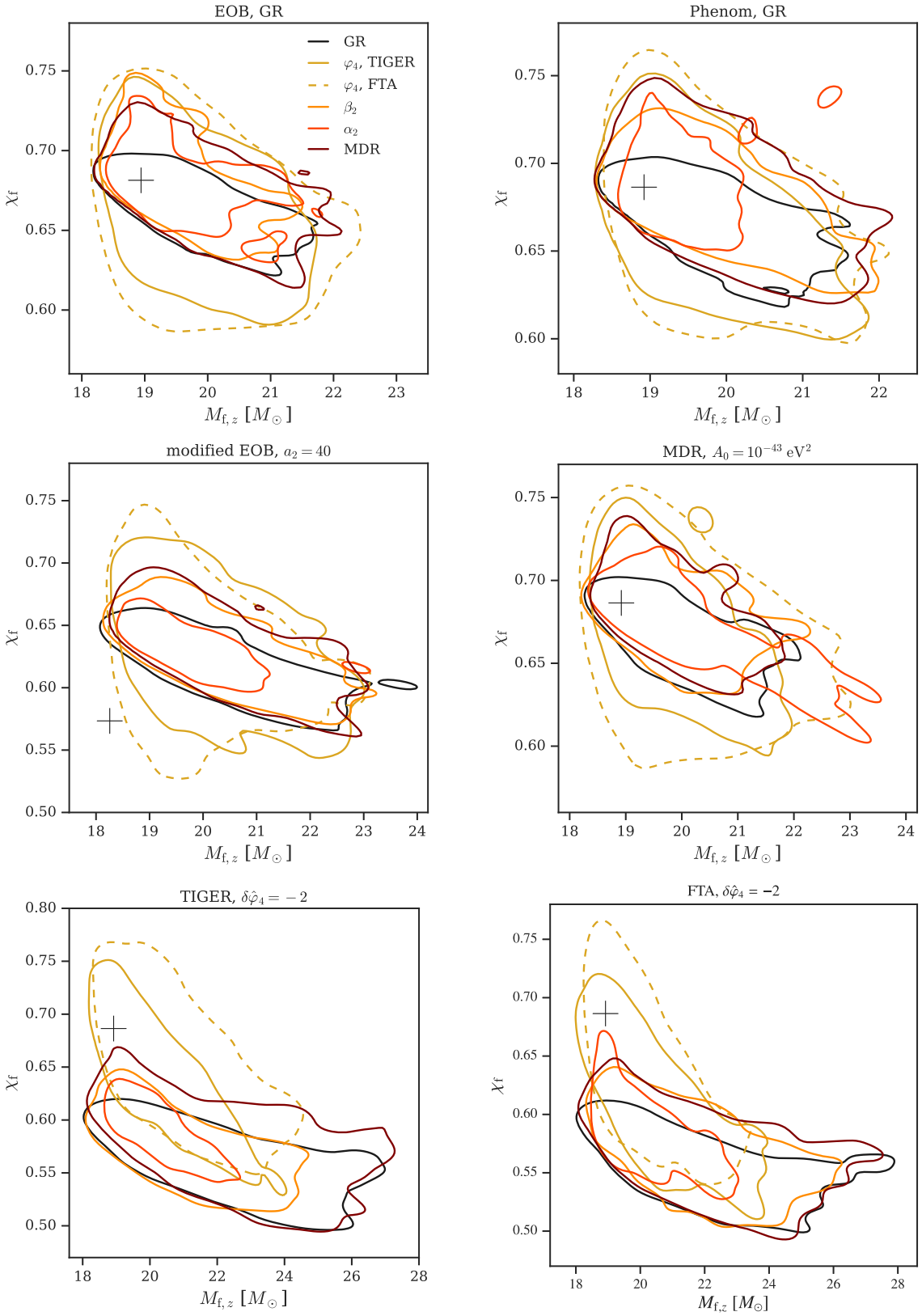

FIG. 12. The analog of Fig. 10 for the GW170608-like simulated observations.

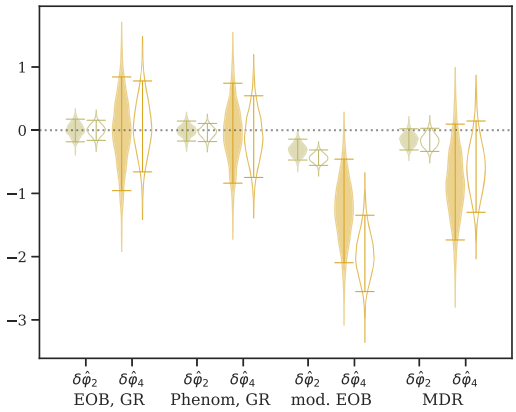

FIG. 13. Comparing the results of the TIGER and FTA φ_2 and φ_4 analyses on the GW150914-like simulated observations with GR waveforms and the modified EOB and massive graviton (MDR) waveforms with the larger GR violations. The FTA results are shown as unfilled violins.

modified EOB simulated observation). Additionally, the posteriors from many of the different tests are disjoint for all of the GR violating cases except for the massive graviton simulated observation.

For the MDR simulated observation, the IMR consistency postinspiral and MDR analyses are the only ones to recover the true values of the final mass and spin in the 90% credible region. These and the IMR consistency inspiral analyses are also the only ones to recover the true values of the individual (redshifted) masses in the 90% credible region, though most analyses recover the true value of the redshifted chirp mass ($M_z\eta^{3/5}$) in the 90% credible region, except for the TIGER β_2 case, which recovers it at the 96% credible level. All the analyses besides the IMR consistency and MDR tests recover unequal masses, with a mass ratio $\lesssim$0.5, and a positive effective spin, generally $\chi_{\rm eff} \gtrsim 0.4$, except for the TIGER β_2 analysis, for which $\chi_{\rm eff} \gtrsim 0.2$. The IMR consistency inspiral analysis also recovers $\chi_{\rm eff} \gtrsim 0.4$ and prefers unequal masses, even though there is support for equal masses. This preference for unequal masses and positive effective spins is not surprising, since both of these act to extend the inspiral, and the signal is stretched out in time by the propagation effect, as illustrated in Fig. 1.

TABLE III. The analog of Table II for the TIGER and FTA φ_2 analyses of the GW150914-like simulated observations with the GR waveforms as well as the modified EOB and massive graviton (MDR) waveforms with the larger GR deviation.

Simulated observation	φ_2, TIGER		φ_2, FTA	
	$Q_{\rm GR}$ (%)	$\delta\hat{\varphi}_2$	$Q_{\rm GR}$ (%)	$\delta\hat{\varphi}_2$
EOB, GR	50	$0.0^{+0.2}_{-0.2}$	49	$0.0^{+0.2}_{-0.2}$
Phenom, GR	54	$0.0^{+0.1}_{-0.2}$	64	$0.0^{+0.1}_{-0.2}$
modified EOB, >	**100**	$-0.3^{+0.2}_{-0.2}$	**100**	$-0.4^{+0.1}_{-0.2}$
MDR, >	93	$-0.1^{+0.1}_{-0.2}$	92	$-0.2^{+0.2}_{-0.1}$

2. Smaller GR deviations

For the GW150914-like simulated observations with smaller GR deviations, we find that none of the tests recover the GR deviations in the modified EOB and massive graviton simulated observations above the 90% credible level (when rounded), though the TIGER α_2 and MDR analyses in the modified EOB case find that GR is excluded at the 88% and 90% credible level, respectively. The massive graviton simulated observations only have GR excluded at most at the 80% credible level, with the MDR analysis. For the TIGER and FTA simulated observation, several tests find GR to be excluded at the 90% credible level or higher, even at slightly greater than 4σ up to slightly greater than 4.5σ for the TIGER and FTA φ_4 tests. This is not surprising, given that the frequency-domain dephasings are significantly larger for the TIGER and FTA cases than for the modified EOB and massive graviton cases, as illustrated in Fig. 3. What is interesting is that the MDR analysis finds GR to be excluded at the 99% credible level for the TIGER simulated observation, even though it only excluded GR at the 42% credible level for the TIGER simulated observation with the larger deviation from GR. This is likely due to the larger GR deviation leading to a very short signal that is easier to fit with a GR template.

For the modified EOB simulated observation, the TIGER and FTA φ_4 analyses not only do not find a deviation from GR but have no support at the true value of $\delta\hat{\varphi}_4 \simeq -2$, showing again that the constraints one obtains from such analyses cannot be straightforwardly interpreted as constraints on PN parameters.

We find the residuals test to be insensitive to all these smaller modifications of GR, returning SNR distributions that are almost identical to those for the GR simulated observations, likely because the GR analysis recovers almost all of the SNR, as seen in Fig. 8. The reconstruction

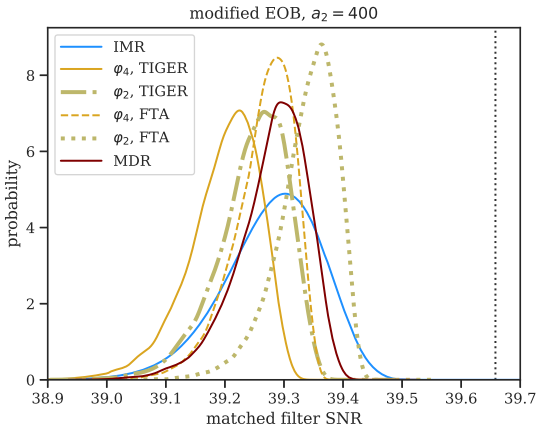

FIG. 14. The analog of Fig. 7 for the TIGER and FTA φ_2 analyses of the GW150914-like modified EOB simulated observation with the larger GR violations. We only show the IMR consistency, TIGER and FTA inspiral, and MDR results to give a cleaner plot.

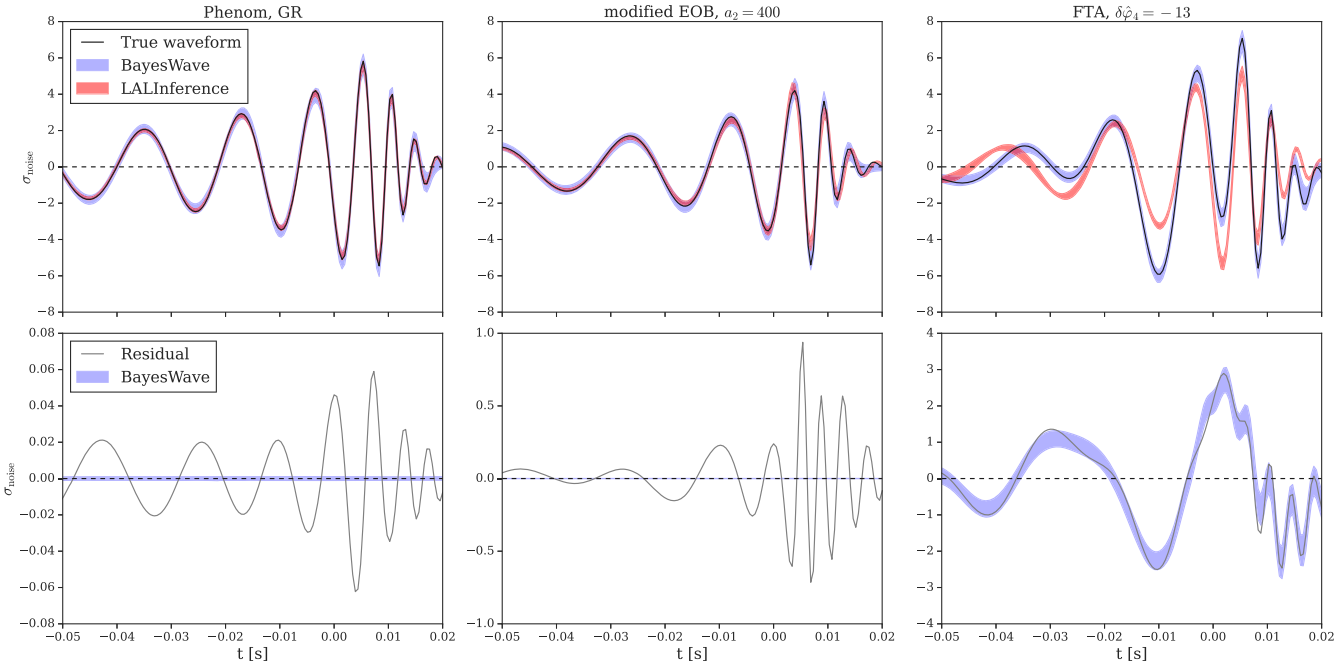

FIG. 15. The waveform reconstruction and residuals analysis results on three cases of the GW150914-like simulated observations: the Phenom GR case and the modified EOB and FTA cases with larger GR deviations. All quantities are shown here as they would appear in the LIGO Livingston detector. The top panels show the true waveform, BAYESWAVE 90% CIs, and the LALInference 90% CI. The bottom panels show the residual data obtained by subtracting the maximum likelihood waveform obtained by the LALInference GR analysis, and the 90% CI obtained by analyzing the residual data using BAYESWAVE. The horizontal axis gives the time from the peak of the waveform, and vertical axis gives the strain amplitude whitened using a filter given by the inverse amplitude spectral density of the detector noise. The whitened strain is measured in units of the standard deviation of the noise, $\sigma_{\rm noise}$. Note that the disagreement between the LALInference reconstruction and the simulated waveform increases from left to right consistent with the increase of the deviation of the waveform morphology from GR from left to right. This is seen clearly in the bottom panels where the average amplitude of the residual time series grows approximately by an order of magnitude in each plot.

comparison finds shifts to larger mismatches for the modified EOB, TIGER, and FTA cases, compared to the GR cases, though not for the massive graviton case. However, there is still considerable overlap of the posteriors in the TIGER and FTA cases. The largest shift is seen for the modified EOB case, and even there the posteriors have some overlap, unlike the disjoint posteriors found in the case with the larger GR deviation. Interestingly, the reconstruction recovers less SNR in the modified EOB case than the median from the GR analysis, as seen in Fig. 8. This is presumably because the signal is more spread out with this smaller GR deviation than in the case of the large GR deviation, where the reconstruction found more SNR than the GR analysis did (see Fig. 7).

The final mass and spin recovery is shown in Fig. 11. We find the same general pattern as before, with the true values for these quantities lying inside the 90% credible regions for the IMR consistency test postinspiral analysis and for the test associated with the waveform in the TIGER and FTA cases. The true values fall just outside of the 90% credible region in the massive graviton case with the MDR analysis, discussed further below. The same patterns in the recovery of the mass ratio and effective spin for the modified EOB and massive graviton simulated observations noted above for the larger GR violations are still present here, just with reduced amplitude. That is, the recoveries of the modified EOB simulated observation prefer close to equal masses and negative effective spins to give a shorter waveform and smaller final spin, while the recoveries of the massive graviton simulated observation prefer unequal masses and positive effective spins, to give a longer waveform. In fact, this preference is even seen in the MDR $A_0 > 0$ recovery of the massive graviton simulated observation and likely explains the bias seen in the final mass and spin noted above.

B. GW170608-like cases

For the GW170608-like cases, we do not consider the IMR consistency test, since it is not applicable to these low-mass, moderate-SNR systems, and is thus not applied to GW170608 in [7,10,11]. The BAYESWAVE analyses are also not as well suited to these more spread-out signals as to the shorter GW150914-like signals considered previously, but we show their results anyway, for comparison, since these analyses are applied to GW170608 itself in [7,78].

In the GW170608-like cases, we find that the tests only identify the GR deviations at or above the 90% credible level in the TIGER case, and even there this is only for the TIGER and FTA φ_4 tests and the TIGER β_2 test. However, in the FTA case the FTA analysis finds a GR deviation at the 88% credible level. The most significant GR deviation for the modified EOB case is again found by the MDR analysis, though this time only at the 72% credible level. The TIGER and FTA analyses also find that the true value of the deviation parameter ($\delta\hat{\varphi}_4 \simeq -2$) is outside the 90% CI, though it is closer here than in the GW150914-like cases. The reconstructions analysis finds a distribution of mismatches for the modified EOB case that is shifted to larger values than for the GR cases (and the other non-GR cases), though the distributions still overlap.

In the massive graviton case, not even the MDR analysis finds a significant deviation from GR. In fact, the true value of $\tilde{A}_0 = 10$ is well outside of the 90% CI. This is due largely to a bias in the recovery of the distance, since the inferred distance determines how one converts the observed dephasing into a bound on A_0. This bias on the distance comes from the distance-inclination degeneracy, where the distance and inclination angle both peak at significantly larger values than the true ones. See, e.g., Fig. 9 in [90] for an example of this bias for a simulated binary black hole observation in Gaussian noise and Fig. 1 in [135] for an example for a simulated binary neutron star observation with zero noise. However, the bias we find is a bit more extreme than in those cases, with a median and 90% CI for the distance of 591^{+82}_{-168} Mpc for the GR analysis of the Phenom GR simulated observation, compared to the true value of 364 Mpc. We find this bias in all of our analyses of the GW170608-like cases. In particular, the MDR recovery with both signs of A_0 gives very similar results for the distance median and 90% CI to the analysis of the Phenom GR simulated observation.

If one uses the true values of the distance and redshift to obtain the posterior on A_0 from the posterior on $\lambda_{A,\mathrm{eff}}$, which is the parameter that directly enters the phase and is sampled on [see, e.g., Eq. (2) in Ref. [7]], and scales $\lambda_{A,\mathrm{eff}}$ by $(D_L/D_L^{\mathrm{true}})^2$ so that the dephasing is unchanged, then one obtains a median and 90% CI for $\tilde{A}_0$ of $1.0^{+8.8}_{-3.9}$, so the true value of $\tilde{A}_0 = 10$ is much closer to being included. There are no noticeable biases in the other parameters, though all the analyses of the massive graviton case favor unequal masses and a positive effective spin, as was found for the other massive graviton cases. This preference may explain the remaining bias in the recovery of A_0.

As illustrated in Fig. 9, the recovered SNR is not significantly different between the different tests except in a few cases. Two of these cases are the TIGER and FTA simulated observations, with their somewhat larger GR deviations. The other cases are the TIGER α_2 test for all simulated observations, with its very broad posteriors on the testing parameter. These cases all have a broader posterior on the SNR, as well, extending to lower values. The recovery of the final mass and spin is shown in Fig. 12. None of the tests find the true values for the modified EOB case in their 90% credible regions and all the analyses prefer unequal masses and a negative effective spin, as we found for the other modified EOB cases. The final mass and spin are recovered in all the 90% credible regions in the massive graviton case, but just for both the TIGER and FTA tests in the TIGER and FTA cases.

V. SUMMARY AND CONCLUSIONS

We have studied how a selection of standard tests of GR that are regularly applied to LIGO-Virgo observations of binary black holes respond to a variety of phenomenological deviations from GR. Specifically, we considered the residuals test, IMR consistency test, TIGER and FTA parametrized tests, and the MDR test. We also considered how well the unmodeled reconstructions of the waveforms agree with the GR waveforms that are found to describe the signal well. The non-GR waveforms we considered are the ones with phenomenological deviations in post-Newtonian coefficients used in the TIGER and FTA tests, as well as the propagation effects from a massive graviton, and a self-consistent modification of the binary's energy flux in the EOB framework. For all of these waveforms, we considered a GW150914-like system with larger and smaller GR deviations and a GW170608-like system with smaller GR deviations. We also considered the GR analogs of the non-GR waveform models considered.

For the GW150914-like case with larger deviations of GR, we found that the deviations from GR are detected at a high credible level by most of the tests considered. However, even for these large deviations, some tests find consistency with GR at the 90% credible level. In particular, in the massive graviton case with the larger graviton mass, only the IMR consistency test, TIGER β_2, and MDR analyses exclude GR at the 90% credible level (and just barely for the IMR consistency test). However, all other cases with large GR deviations are identified as deviations from GR at the 90% credible level or greater by at least five tests. Indeed, many of the larger GR deviations are identified as such at very high credible levels, greater than a Gaussian 5σ. (These very high credible levels are likely because our simulated observations do not contain noise.) For the GW150914-like smaller GR deviations, the number of tests that find a significant GR deviation decreases considerably. Most notably, none of the tests identify the massive graviton case as a GR violation above the 80% credible level and the modified EOB case is only (just) identified as a GR violation at the 90% credible level by the MDR analysis. However, the TIGER and FTA modifications are identified as GR violations at the 90%

credible level or greater by all but two of the tests considered.

For the GW170608-like case, with its smaller SNR, we found that only the TIGER case is identified as a GR deviation at the 90% credible level (by the TIGER and FTA tests), though the FTA analysis almost identifies the FTA case as a GR deviation at the 90% credible level and the MDR analysis identifies the modified EOB waveform as a GR violation at the 72% credible level. These are the only cases that are identified as GR violations at such high credible levels.

One does not always find that the tests one expects to detect a given GR violation strongly are actually effective in doing so. Conversely, one finds that tests that one might not expect to be effective in detecting a given GR violation detect it strongly. The most striking example of both of these is likely the GW150914-like modified EOB waveform with the smaller GR deviation. Here one might expect that the TIGER and FTA tests that look for deviations in the 2PN phase coefficient would find significant deviations from GR, since the leading deviation in the inspiral phase in the modified EOB waveform is at 2PN. However, this is not the case: Both of these tests find excellent consistency with GR in this case, while the MDR analysis recovers a deviation from GR at the 90% credible level (albeit just barely). For another case where the TIGER and FTA tests of inspiral PN coefficients do not recover the deviation from GR as strongly as one might expect, in the GW150914-like massive graviton case with the larger GR deviation, where the leading order of the deviation in the inspiral is at 1PN, the TIGER β_2 intermediate coefficient test finds a GR deviation at the 98% credible level, while the TIGER and FTA 1PN analyses only find a deviation at about the 85% credible level.

In fact, for the modified EOB waveform, the 2PN TIGER and FTA analyses do not even recover the true value of the deviation parameter within the 90% CI. This is particularly true for the GW150914-like cases, though in the case with the larger GR deviation, the TIGER and FTA analyses do find a strong deviation from GR, even though they underestimate the size of the deviation parameter by almost an order of magnitude. In the GW170608-like case, the true value of the deviation parameter is slightly closer to the boundary of the 90% credible region than in the GW150914-like case with the smaller GR deviation, but still outside it. The 1PN TIGER and FTA analyses of the GW150914-like massive graviton case with the larger massive graviton mass (the only massive graviton case we analyze with the 1PN analyses) also recover significantly smaller deviations than the true value.

The fact that the TIGER and FTA analyses do not recover the true value of the modified PN parameter is not surprising. These analyses are designed to detect deviations from GR, not to measure individual PN coefficients: They only modify one PN coefficient at a time and include the post-inspiral part of the signal in the analysis without attempting to account for the expected modifications to this part of the waveform in modified theories. This means that analyses that interpret the TIGER and FTA results as constraints on PN parameters, e.g., [22,24], may be obtaining apparent constraints on modified theories that are significantly more stringent than actually allowed by the data. Additionally, this suggests that it would be a good idea to apply similar checks to the method for modifying the PN coefficients in [27], where the frequency domain dephasing is applied to the entire signal. Given the results here, it seems likely that this method will also underestimate the size of a potential GR deviation, invalidating the constraints on alternative theories presented there. Developing a method to constrain deviations from PN coefficients accurately in as generic a situation as possible would be a very worthwhile endeavor. The method in [24] that restricts to the low-frequency portion of the signal is a possible way to proceed, though it would still need to be validated with these sorts of tests. In particular, it seems unlikely that the current setting of the IMRPhenomD value for the end of the inspiral for the high-frequency cutoff is the optimal choice. We provide the frame files for our simulated observations [131] so they can be used to perform such checks.

One also finds that the residuals test is not very sensitive to most of the deviations from GR considered here. It only excludes GR for the extreme deviations from GR in the GW150914-like TIGER and FTA cases with the larger GR violations, where the waveforms do not look at all like those from binary black hole coalescences in GR (see Fig. 1). Thus, while the residuals test seems like a promising way to identify deviations from GR (or more generally from the quasicircular binary black hole hypothesis) without making assumptions about the exact nature of the deviations, it is likely only effective in detecting extreme deviations from GR, at least for the relatively moderate SNRs that one expects for most detections by current and near-future detectors.

The comparison of unmodeled and GR reconstructions appears to be more effective at identifying deviations from GR than the residuals test: The distribution of mismatches between the two reconstructions is well separated from the distribution for the GR waveform in several cases where the residuals test does not identify any deviation from GR, notably for the GW150914-like modified EOB waveform with the larger GR deviation. However, in our analysis we are only comparing the mismatches between reconstructions to a single GR case and with no noise. It is likely that the expected distribution of mismatches in the GR case would broaden considerably when considering a larger range of GR waveforms and detector noise, considerably

weakening these results. Nevertheless, it is likely worth pursuing the reconstruction comparison as a test of GR for high-mass binary black hole signals. It will, however, not be applicable to low-mass signals, like the GW170608-like cases we consider, where the power is spread out over about a second or more, making it difficult for the unmodeled reconstructions to recover the waveform accurately.

Finally, we found that the final mass and spin distributions recovered by the different tests have disjoint 90% credible regions for many of the tests with larger GR deviations. This suggests that it might be worthwhile to develop a "meta IMR consistency test" by comparing the recovery of the final mass and spin (or other parameters) between different tests.

Of course, this is still quite a preliminary study, and there is much more to do to assess the relation between different tests of GR on gravitational wave data. For instance, it is important to consider the effects on tests of GR of missing physics in the waveform models, e.g., higher modes, for which there are initial studies in Refs. [136,137], as well as eccentricity (for which there are fairly well developed numerical relativity calculations for binary black holes and some waveform models that reproduce these results reasonably well in certain portions of the parameter space, e.g., Refs. [138–151]). Other important physical effects to consider are those from gravitational lensing (e.g., the effects calculated in Ref. [152] in the geometrical optics regime as well as wave optics effects [153,154]), and the presence of a third body (e.g., the calculations in [155–162]) and other environmental effects (e.g., from gas or dark matter) [158,163].

Similarly, one should consider waveforms from binaries of black hole mimickers (see [63,64] for some simple checks using rescaled binary neutron star and black hole–neutron star waveforms, respectively, and [164] for a toy model for such waveforms). Finally, one needs to assess the effects of systematic errors in the baseline GR waveform models, which could plausibly start to affect current combined constraints, as discussed in [165], and will definitely be important even for loud individual events in future detectors, as discussed in, e.g., Refs. [166,167].

One will also want to include more tests in future studies and use waveforms from various alternative theories, once they are computed with sufficient accuracy (there are also constructions of self-consistent waveforms based on analytical knowledge of modified theories [168], also used in [169], that could be useful in these sorts of studies before full numerical waveforms are available). It is also important to consider the effects of detector noise and calibration as well as systematics in the GR waveform models; see [170] for a study of the effects of transient non-Gaussian noise features ("glitches") and their removal on TIGER. However, the most important study will likely be considering populations of signals to determine how well the tests perform when combining together multiple observations to potentially detect smaller deviations from GR, e.g., using the method in [171]. Here it will be particularly important to include the effects of spins and higher modes in the simulated observations, which were not included in this initial study.

Nevertheless, this study already indicates that one will require quite high SNRs, above the SNRs of ∼50 we considered here in the GW150914-like case, to be able to detect some moderate deviations from GR in individual events with the tests we consider here. This strongly motivates the need for improvements in gravitational wave detectors, particularly third generation ground-based detectors [13–15], to provide the much larger SNRs that will allow one to distinguish relatively small deviations from GR. Additionally, improvements in the design of tests of GR and methods for combining together multiple observations will also be necessary to fully exploit current and future gravitational wave detector data to test GR.

ACKNOWLEDGMENTS

We wish to thank all the LIGO-Virgo-KAGRA testing GR group members who implemented these tests in publicly available code. Additionally, we thank Anuradha Samajdar for assistance with the MDR test, initial work on this project, and a careful reading of the paper, Archisman Ghosh for the code used to create the frame files to analyze, Noah Sennett and Michalis Agathos for the FTA reweighting script, Parameswaran Ajith for initial discussions, and Chris Van Den Broeck and B. S. Sathyaprakash for useful comments. N. K. J.-M. acknowledges support from STFC Consolidator Grant No. ST/L000636/1. S. G. gratefully acknowledges support from National Science Foundation (NSF) grant No. PHY-1809572. M. S. acknowledges support from the Infosys Foundation, the Swarnajayanti fellowship grant No. DST/SJF/PSA-01/2017-18, and NSF grants No. PHY-00090754, No. PHY-1806630, and No. PHY-2010970. N. V. K. acknowledges support from the Max Planck Society's Independent Research Group Grant. J. A. C. acknowledges support from NSF grants No. PHY-1700765 and No. PHY-1764464. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by NSF Grants No. PHY-0757058 and No. PHY-0823459 as well as by the Open Science Grid [172,173], which is supported by the NSF award No. 2030508. Additional computations were performed on the clusters Alice at the International Centre for Theoretical Sciences, Tata Institute of Fundamental Research and Hypatia at the Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Potsdam-Golm. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the NSF. This research has

made use of data obtained from the Gravitational Wave Open Science Center (www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO is funded by the US NSF. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. We used the following software in this study: BAYESWAVE [92,93], LALSuite [91], MATPLOTLIB [174], NUMPY [175], PESummary [176], PyCBC [177], SCIPY [178], and SEABORN [179]. This is LIGO document P2100322.

APPENDIX: 2D IMR CONSISTENCY PLOTS

Here we give the 2D $\Delta M_f/\bar{M}_f$, $\Delta\chi_f/\bar{\chi}_f$ joint probability distributions for the IMR consistency test, for comparison with analogous plots shown for analyses of gravitational wave detections in Refs. [10,11]. (The 2D plots in the methods papers [83,84] and earlier LIGO-Virgo papers [1,3,7,180] are not exactly comparable, since they do not use flat priors in $\Delta M_f/\bar{M}_f$ and $\Delta\chi_f/\bar{\chi}_f$, and [1,83] also use a different normalization.) We show two sets of results in Fig. 16. We first show the results corresponding to the results shown in Sec. IV, which infer $f_{\rm cut}$ from the full IMR GR analysis of each of the simulated observations, as

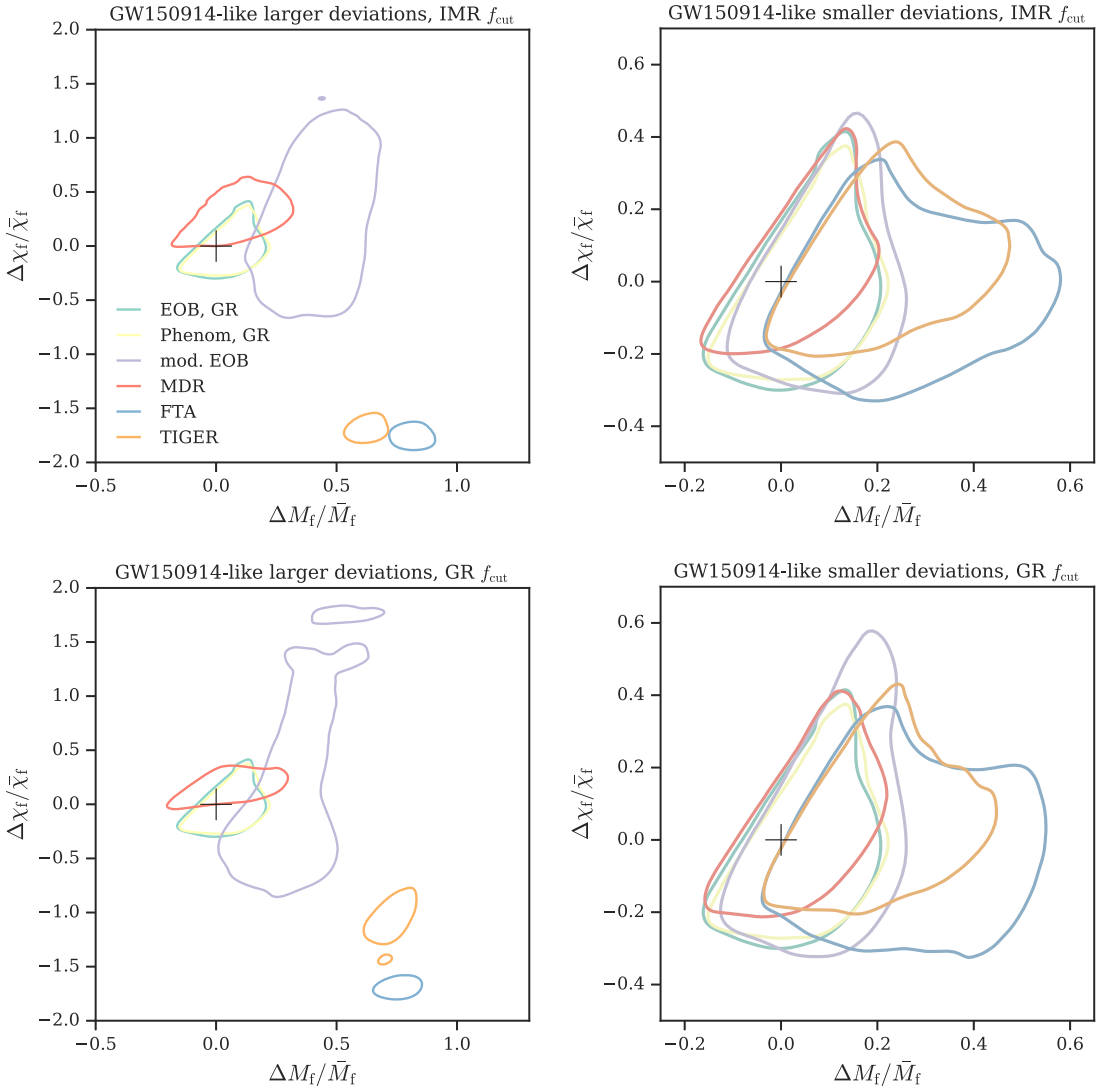

FIG. 16. The 90% credible regions of the joint probability distributions of $\Delta M_f/\bar{M}_f$, $\Delta\chi_f/\bar{\chi}_f$ for the IMR consistency test applied to the GW150914-like simulated observations. The left (right) column shows the results with larger (smaller) GR deviations. The top row shows the results with the test's cutoff frequency obtained using the GR analysis of the simulated observation being analyzed (given in Table IV), while the bottom row fixes these to the values from each simulated observation's corresponding GR case. The results for the GR simulated observations are the same in all four panels.

discussed in Sec. II C. Then, for comparison, we show the results obtained with the same $f_{\rm cut}$ as the corresponding GR simulated observation (i.e., $f_{\rm cut} = 129$ Hz for the modified EOB observations and $f_{\rm cut} = 131$ Hz for all the others).

As expected, the difference between the results with the IMR $f_{\rm cut}$ and GR simulated observation $f_{\rm cut}$ is largest for the modified EOB case with the larger GR deviation, since this is the case with the largest difference between the two cutoff frequencies. Surprisingly, the 2 Hz difference in the TIGER case with the larger GR deviation leads to disjoint probability distributions, due to the extreme GR deviation in this case. In all other cases, the differences with different $f_{\rm cut}$ values are not so significant, with substantial overlap of the probability distributions, even in the MDR case with the larger GR deviation, which has almost as large a difference in $f_{\rm cut}$ as the modified EOB case with the larger GR deviation, though in the opposite direction.

TABLE IV. Cutoff frequencies $f_{\rm cut}$ inferred from the GR analysis of each of the GW150914-like simulated observations, rounded to the nearest Hz. As in Table II, > denotes the case with the larger GR deviation and < the one with the smaller deviation. The results are given in two groups, first the EOB GR result and the results for the modified EOB waveforms and then the Phenom GR result and the results for the Phenom-based non-GR waveforms.

	$f_{\rm cut}$ (Hz)		
Simulated observation	>		<
EOB, GR		129	
Modified EOB	92		122
Phenom, GR		131	
MDR	164		137
TIGER	129		125
FTA	117		121

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 221101 (2016).
[2] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **6**, 041015 (2016).
[3] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **118**, 221101 (2017).
[4] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **119**, 141101 (2017).
[5] B. P. Abbott *et al.* (LIGO Scientific, Virgo, Fermi-GBM, and INTEGRAL Collaborations), Astrophys. J. Lett. **848**, L13 (2017).
[6] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **123**, 011102 (2019).
[7] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **100**, 104036 (2019).
[8] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **125**, 101102 (2020).
[9] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. Lett. **900**, L13 (2020).
[10] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **103**, 122002 (2021).
[11] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), arXiv:2112.06861.
[12] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, and Virgo Collaborations), Living Rev. Relativity **23**, 3 (2020), noise curves available from https://git.ligo.org/lscsoft/lalsuite/tree/master/lalsimulation/lib.
[13] S. Hild *et al.*, Classical Quantum Gravity **28**, 094013 (2011).
[14] D. Reitze *et al.*, Bull. Am. Astron. Soc. **51**, 35 (2019), https://ui.adsabs.harvard.edu/abs/2019BAAS...51g..35R/abstract.
[15] E. D. Hall *et al.*, Phys. Rev. D **103**, 122004 (2021).
[16] M. Maggiore *et al.*, J. Cosmol. Astropart. Phys. 03 (2020) 050.
[17] J. Baker *et al.*, arXiv:1907.06482.
[18] P. Amaro-Seoane *et al.*, arXiv:1702.00786.
[19] S. E. Perkins, N. Yunes, and E. Berti, Phys. Rev. D **103**, 044024 (2021).
[20] N. Yunes, K. Yagi, and F. Pretorius, Phys. Rev. D **94**, 084002 (2016).
[21] S. Tahura and K. Yagi, Phys. Rev. D **98**, 084042 (2018); **101**, 109902(E) (2020).
[22] R. Nair, S. Perkins, H. O. Silva, and N. Yunes, Phys. Rev. Lett. **123**, 191101 (2019); **124**, 169904(E) (2020).
[23] S. Tahura, K. Yagi, and Z. Carson, Phys. Rev. D **100**, 104001 (2019).
[24] H.-T. Wang, S.-P. Tang, P.-C. Li, M.-Z. Han, and Y.-Z. Fan, Phys. Rev. D **104**, 024015 (2021).
[25] A. J. K. Chua and M. Vallisneri, arXiv:2006.08918.
[26] N. Sennett, R. Brito, A. Buonanno, V. Gorbenko, and L. Senatore, Phys. Rev. D **102**, 044056 (2020).
[27] S. E. Perkins, R. Nair, H. O. Silva, and N. Yunes, Phys. Rev. D **104**, 024060 (2021).
[28] E. W. Hirschmann, L. Lehner, S. L. Liebling, and C. Palenzuela, Phys. Rev. D **97**, 064032 (2018).
[29] H. Witek, L. Gualtieri, P. Pani, and T. P. Sotiriou, Phys. Rev. D **99**, 064035 (2019).
[30] M. Okounkova, L. C. Stein, J. Moxon, M. A. Scheel, and S. A. Teukolsky, Phys. Rev. D **101**, 104016 (2020).
[31] M. Okounkova, Phys. Rev. D **102**, 084046 (2020).
[32] W. E. East and J. L. Ripley, Phys. Rev. D **103**, 044040 (2021).
[33] H. O. Silva, H. Witek, M. Elley, and N. Yunes, Phys. Rev. Lett. **127**, 031101 (2021).
[34] W. E. East and J. L. Ripley, Phys. Rev. Lett. **127**, 101102 (2021).
[35] E. Barausse, C. Palenzuela, M. Ponce, and L. Lehner, Phys. Rev. D **87**, 081506(R) (2013).

[36] M. Shibata, K. Taniguchi, H. Okawa, and A. Buonanno, Phys. Rev. D **89**, 084005 (2014).
[37] M. Bezares, R. Aguilera-Miret, L. ter Haar, M. Crisostomi, C. Palenzuela, and E. Barausse, arXiv:2107.05648.
[38] N. Sennett, S. Marsat, and A. Buonanno, Phys. Rev. D **94**, 084003 (2016).
[39] L. Bernard, Phys. Rev. D **98**, 044004 (2018).
[40] L. Bernard, Phys. Rev. D **99**, 044047 (2019).
[41] L. Bernard, Phys. Rev. D **101**, 021501(R) (2020).
[42] M. Khalil, N. Sennett, J. Steinhoff, J. Vines, and A. Buonanno, Phys. Rev. D **98**, 104010 (2018).
[43] F.-L. Julié, J. Cosmol. Astropart. Phys. 10 (2018) 033.
[44] F.-L. Julié and E. Berti, Phys. Rev. D **100**, 104061 (2019).
[45] M. Accettulli Huber, A. Brandhuber, S. De Angelis, and G. Travaglini, Phys. Rev. D **103**, 045015 (2021).
[46] B. Shiralilou, T. Hinderer, S. M. Nissanke, N. Ortiz, and H. Witek, Phys. Rev. D **103**, L121503 (2021).
[47] B. Shiralilou, T. Hinderer, S. Nissanke, N. Ortiz, and H. Witek, arXiv:2105.13972.
[48] P. Brax, A.-C. Davis, S. Melville, and L. K. Wong, J. Cosmol. Astropart. Phys. 10 (2021) 075.
[49] E. Battista and V. De Falco, Phys. Rev. D **104**, 084067 (2021).
[50] G. Bozzola and V. Paschalidis, Phys. Rev. Lett. **126**, 041103 (2021).
[51] G. Bozzola and V. Paschalidis, Phys. Rev. D **104**, 044004 (2021).
[52] S. L. Liebling and C. Palenzuela, Living Rev. Relativity **20**, 5 (2017).
[53] V. Cardoso, S. Hopper, C. F. B. Macedo, C. Palenzuela, and P. Pani, Phys. Rev. D **94**, 084031 (2016).
[54] M. Bezares, C. Palenzuela, and C. Bona, Phys. Rev. D **95**, 124005 (2017).
[55] C. Palenzuela, P. Pani, M. Bezares, V. Cardoso, L. Lehner, and S. Liebling, Phys. Rev. D **96**, 104058 (2017).
[56] T. Helfer, E. A. Lim, M. A. G. Garcia, and M. A. Amin, Phys. Rev. D **99**, 044046 (2019).
[57] N. Sanchis-Gual, C. Herdeiro, J. A. Font, E. Radu, and F. Di Giovanni, Phys. Rev. D **99**, 024017 (2019).
[58] M. Bezares and C. Palenzuela, Classical Quantum Gravity **35**, 234002 (2018).
[59] T. Helfer, U. Sperhake, R. Croft, M. Radia, B.-X. Ge, and E. A. Lim, arXiv:2108.11995.
[60] N. V. Krishnendu, K. G. Arun, and C. K. Mishra, Phys. Rev. Lett. **119**, 091101 (2017).
[61] N. V. Krishnendu, M. Saleem, A. Samajdar, K. G. Arun, W. Del Pozzo, and C. K. Mishra, Phys. Rev. D **100**, 104019 (2019).
[62] S. Dhanpal, A. Ghosh, A. K. Mehta, P. Ajith, and B. S. Sathyaprakash, Phys. Rev. D **99**, 104056 (2019).
[63] T. Islam, A. K. Mehta, A. Ghosh, V. Varma, P. Ajith, and B. S. Sathyaprakash, Phys. Rev. D **101**, 024032 (2020).
[64] N. K. Johnson-McDaniel, A. Mukherjee, R. Kashyap, P. Ajith, W. Del Pozzo, and S. Vitale, Phys. Rev. D **102**, 123010 (2020).
[65] S. Kastha, A. Gupta, K. G. Arun, B. S. Sathyaprakash, and C. Van Den Broeck, Phys. Rev. D **98**, 124033 (2018).
[66] S. Kastha, A. Gupta, K. G. Arun, B. S. Sathyaprakash, and C. Van Den Broeck, Phys. Rev. D **100**, 044007 (2019).
[67] G. Carullo, G. Riemenschneider, K. W. Tsang, A. Nagar, and W. Del Pozzo, Classical Quantum Gravity **36**, 105009 (2019).
[68] G. Carullo, W. Del Pozzo, and J. Veitch, Phys. Rev. D **99**, 123029 (2019); **100**, 089903(E) (2019).
[69] A. Maselli, P. Pani, L. Gualtieri, and E. Berti, Phys. Rev. D **101**, 024043 (2020).
[70] G. Carullo, Phys. Rev. D **103**, 124043 (2021).
[71] A. Ghosh, R. Brito, and A. Buonanno, Phys. Rev. D **103**, 124041 (2021).
[72] Y. Asali, P. T. H. Pang, A. Samajdar, and C. Van Den Broeck, Phys. Rev. D **102**, 024016 (2020).
[73] C.-J. Haster, arXiv:2005.05472.
[74] C. D. Capano and A. H. Nitz, Phys. Rev. D **102**, 124070 (2020).
[75] B. Edelman, F. J. Rivera-Paleo, J. D. Merritt, B. Farr, Z. Doctor, J. Brink, W. M. Farr, J. Gair, J. S. Key, J. McIver, and A. B. Nielsen, Phys. Rev. D **103**, 042004 (2021).
[76] D. Psaltis, C. Talbot, E. Payne, and I. Mandel, Phys. Rev. D **103**, 104036 (2021).
[77] S. Bhagwat and C. Pacilio, Phys. Rev. D **104**, 024030 (2021).
[78] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **9**, 031040 (2019).
[79] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **11**, 021053 (2021).
[80] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), arXiv:2111.03606.
[81] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[82] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. Lett. **851**, L35 (2017).
[83] A. Ghosh, A. Ghosh, N. K. Johnson-McDaniel, C. K. Mishra, P. Ajith, W. Del Pozzo, D. A. Nichols, Y. Chen, A. B. Nielsen, C. P. L. Berry, and L. London, Phys. Rev. D **94**, 021101(R) (2016).
[84] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Classical Quantum Gravity **35**, 014002 (2018).
[85] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).
[86] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[87] A. Bohé, M. Hannam, S. Husa, F. Ohme, M. Puerrer, and P. Schmidt, PhenomPv2—Technical Notes for LAL Implementation, Tech. Rep. LIGO-T1500602 (LIGO Project, 2016), https://dcc.ligo.org/LIGO-T1500602/public.
[88] A. Bohé *et al.*, Phys. Rev. D **95**, 044028 (2017).
[89] J. Skilling, AIP Conf. Proc. **735**, 395 (2004).
[90] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[91] LSC Algorithm Library Suite (LALSuite), 10.7935/GT1W-FZ16.
[92] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015).
[93] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, Phys. Rev. D **103**, 044006 (2021).

[94] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 074001 (2015).
[95] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Classical Quantum Gravity **33**, 134001 (2016).
[96] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Classical Quantum Gravity **37**, 055002 (2020).
[97] S. Ghonge, K. Chatziioannou, J. A. Clark, T. Littenberg, M. Millhouse, L. Cadonati, and N. Cornish, Phys. Rev. D **102**, 064056 (2020).
[98] J. M. Bardeen, W. H. Press, and S. A. Teukolsky, Astrophys. J. **178**, 347 (1972).
[99] F. Hofmann, E. Barausse, and L. Rezzolla, Astrophys. J. Lett. **825**, L19 (2016).
[100] J. Healy and C. O. Lousto, Phys. Rev. D **95**, 024037 (2017).
[101] X. Jiménez-Forteza, D. Keitel, S. Husa, M. Hannam, S. Khan, and M. Pürrer, Phys. Rev. D **95**, 064024 (2017).
[102] N. K. Johnson-McDaniel *et al.*, Determining the final spin of a binary black hole system including in-plane spins: Method and checks of accuracy, Tech. Rep. LIGO-T1600168 (LIGO Project, 2016), https://dcc.ligo.org/LIGO-T1600168/public/main.
[103] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Classical Quantum Gravity **23**, L37 (2006).
[104] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Phys. Rev. D **74**, 024006 (2006).
[105] N. Yunes and F. Pretorius, Phys. Rev. D **80**, 122003 (2009).
[106] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, Phys. Rev. D **85**, 082003 (2012).
[107] L. Blanchet, Living Rev. Relativity **17**, 2 (2014).
[108] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, Phys. Rev. D **89**, 082001 (2014).
[109] J. Meidam *et al.*, Phys. Rev. D **97**, 044033 (2018).
[110] A. Gupta, S. Datta, S. Kastha, S. Borhanian, K. G. Arun, and B. S. Sathyaprakash, Phys. Rev. Lett. **125**, 201101 (2020).
[111] S. Datta, A. Gupta, S. Kastha, K. G. Arun, and B. S. Sathyaprakash, Phys. Rev. D **103**, 024036 (2021).
[112] A. A. Shoom, P. K. Gupta, B. Krishnan, A. B. Nielsen, and C. D. Capano, arXiv:2105.02191.
[113] M. Saleem, S. Datta, K. G. Arun, and B. S. Sathyaprakash, arXiv:2110.10147.
[114] S. Mirshekari, N. Yunes, and C. M. Will, Phys. Rev. D **85**, 024041 (2012).
[115] P. A. R. Ade *et al.* (Planck Collaboration), Astron. Astrophys. **594**, A13 (2016).
[116] T. Damour, A. Nagar, and S. Bernuzzi, Phys. Rev. D **87**, 084035 (2013).
[117] IHES EOB code http://eob.ihes.fr; we use the 12.02 version of the code.
[118] A. Nagar *et al.*, Phys. Rev. D **98**, 104052 (2018).
[119] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Phys. Rev. D **98**, 084028 (2018).
[120] A. Nagar, G. Riemenschneider, G. Pratten, P. Rettegno, and F. Messina, Phys. Rev. D **102**, 024077 (2020).
[121] S. Ossokine, A. Buonanno, S. Marsat, R. Cotesta, S. Babak, T. Dietrich, R. Haas, I. Hinder, H. P. Pfeiffer, M. Pürrer, C. J. Woodford, M. Boyle, L. E. Kidder, M. A. Scheel, and B. Szilágyi, Phys. Rev. D **102**, 044055 (2020).
[122] S. Akçay, R. Gamba, and S. Bernuzzi, Phys. Rev. D **103**, 024014 (2021).
[123] R. Gamba, S. Akçay, S. Bernuzzi, and J. Williams, arXiv:2111.03675.
[124] M. Ruiz, M. Alcubierre, D. Núñez, and R. Takahashi, Gen. Relativ. Gravit. **40**, 2467 (2008).
[125] C. Cutler and É. É. Flanagan, Phys. Rev. D **49**, 2658 (1994).
[126] SXS Gravitational Waveform Database, http://www.black-holes.org/waveforms.
[127] M. Boyle *et al.*, Classical Quantum Gravity **36**, 195006 (2019).
[128] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **94**, 064035 (2016).
[129] LIGO Scientific and Virgo Collaborations, GWTC-1, 10.7935/82H3-HH23 (2018).
[130] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).
[131] Frame files for our simulated observations, 10.5281/zenodo.5637360.
[132] S. Nissanke, D. E. Holz, S. A. Hughes, N. Dalal, and J. L. Sievers, Astrophys. J. **725**, 496 (2010).
[133] É. Racine, Phys. Rev. D **78**, 044021 (2008).
[134] L. Santamaría, F. Ohme, P. Ajith, B. Brügmann, N. Dorband, M. Hannam, S. Husa, P. Mösta, D. Pollney, C. Reisswig, E. L. Robinson, J. Seiler, and B. Krishnan, Phys. Rev. D **82**, 064016 (2010).
[135] C. L. Rodriguez, B. Farr, V. Raymond, W. M. Farr, T. B. Littenberg, D. Fazi, and V. Kalogera, Astrophys. J. **784**, 119 (2014).
[136] P. T. H. Pang, J. Calderón Bustillo, Y. Wang, and T. G. F. Li, Phys. Rev. D **98**, 024019 (2018).
[137] T. Islam, arXiv:2111.00111.
[138] I. Hinder, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **98**, 044015 (2018).
[139] E. A. Huerta, R. Haas, S. Habib, A. Gupta, A. Rebei, V. Chavva, D. Johnson, S. Rosofsky, E. Wessel, B. Agarwal, D. Luo, and W. Ren, Phys. Rev. D **100**, 064003 (2019).
[140] A. Ramos-Buades, S. Husa, G. Pratten, H. Estellés, C. García-Quirós, M. Mateu-Lucena, M. Colleoni, and R. Jaume, Phys. Rev. D **101**, 083015 (2020).
[141] X. Liu, Z. Cao, and L. Shao, Phys. Rev. D **101**, 044049 (2020).
[142] D. Chiaramello and A. Nagar, Phys. Rev. D **101**, 101501 (R) (2020).
[143] Z. Chen, E. A. Huerta, J. Adamo, R. Haas, E. O'Shea, P. Kumar, and C. Moore, Phys. Rev. D **103**, 084018 (2021).
[144] V. Gayathri, J. Healy, J. Lange, B. O'Brien, M. Szczepanczyk, I. Bartos, M. Campanelli, S. Klimenko, C. Lousto, and R. O'Shaughnessy, arXiv:2009.05461.
[145] Y. Setyawati and F. Ohme, Phys. Rev. D **103**, 124011 (2021).
[146] T. Islam, V. Varma, J. Lodman, S. E. Field, G. Khanna, M. A. Scheel, H. P. Pfeiffer, D. Gerosa, and L. E. Kidder, Phys. Rev. D **103**, 064022 (2021).
[147] X. Liu, Z. Cao, and Z.-H. Zhu, arXiv:2102.08614.

[148] Q. Yun, W.-B. Han, X. Zhong, and C. A. Benavides-Gallego, Phys. Rev. D **103**, 124053 (2021).
[149] A. Nagar and P. Rettegno, Phys. Rev. D **104**, 104004 (2021).
[150] A. Placidi, S. Albanesi, A. Nagar, M. Orselli, S. Bernuzzi, and G. Grignani, arXiv:2112.05448.
[151] A. Ramos-Buades, A. Buonanno, M. Khalil, and S. Ossokine, arXiv:2112.06952 [Phys. Rev. D (to be published)].
[152] J. M. Ezquiaga, D. E. Holz, W. Hu, M. Lagos, and R. M. Wald, Phys. Rev. D **103**, 064047 (2021).
[153] R. Takahashi and T. Nakamura, Astrophys. J. **595**, 1039 (2003).
[154] G. Pagano, O. A. Hannuksela, and T. G. F. Li, Astron. Astrophys. **643**, A167 (2020).
[155] Y. Meiron, B. Kocsis, and A. Loeb, Astrophys. J. **834**, 200 (2017).
[156] D. J. D'Orazio and A. Loeb, Phys. Rev. D **101**, 083031 (2020).
[157] P. Gupta, H. Suzuki, H. Okawa, and K.-i. Maeda, Phys. Rev. D **101**, 104053 (2020).
[158] A. Toubiana, L. Sberna, A. Caputo, G. Cusin, S. Marsat, K. Jani, S. Babak, E. Barausse, C. Caprini, P. Pani, A. Sesana, and N. Tamanini, Phys. Rev. Lett. **126**, 101105 (2021).
[159] V. Cardoso, F. Duque, and G. Khanna, Phys. Rev. D **103**, L081501 (2021).
[160] R. S. Chandramouli and N. Yunes, arXiv:2107.00741 [Phys. Rev. D (to be published)].
[161] H. Yu, Y. Wang, B. Seymour, and Y. Chen, Phys. Rev. D **104**, 103011 (2021).
[162] L. Gondán and B. Kocsis, arXiv:2110.09540.
[163] E. Barausse, V. Cardoso, and P. Pani, Phys. Rev. D **89**, 104059 (2014).
[164] A. Toubiana, S. Babak, E. Barausse, and L. Lehner, Phys. Rev. D **103**, 064042 (2021).
[165] C. J. Moore, E. Finch, R. Buscicchio, and D. Gerosa, iScience **24**, 102577 (2021).
[166] M. Pürrer and C.-J. Haster, Phys. Rev. Research **2**, 023151 (2020).
[167] D. Ferguson, K. Jani, P. Laguna, and D. Shoemaker, Phys. Rev. D **104**, 044037 (2021).
[168] Z. Carson and K. Yagi, Classical Quantum Gravity **37**, 215007 (2020).
[169] Z. Carson and K. Yagi, Phys. Rev. D **101**, 104030 (2020).
[170] J. Y. L. Kwok, R. K. L. Lo, A. J. Weinstein, and T. G. F. Li, arXiv:2109.07642 [Phys. Rev. D (to be published)].
[171] M. Isi, K. Chatziioannou, and W. M. Farr, Phys. Rev. Lett. **123**, 121101 (2019).
[172] R. Pordes *et al.*, J. Phys. Conf. Ser. **78**, 012057 (2007).
[173] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Wurthwein, in Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering (IEEE Computer Society, Los Alamitos, 2009), Vol. 2, pp. 428–432.
[174] J. D. Hunter, Comput. Sci. Eng. **9**, 90 (2007).
[175] C. R. Harris *et al.*, Nature (London) **585**, 357 (2020).
[176] C. Hoy and V. Raymond, SoftwareX **15**, 100765 (2021).
[177] A. H. Nitz *et al.*, PyCBC software, 10.5281/zenodo.596388.
[178] P. Virtanen *et al.*, Nat. Methods **17**, 261 (2020).
[179] M. Waskom and the seaborn development team, SEABORN package, 10.5281/zenodo.592845 (2020).
[180] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **102**, 043015 (2020).